The Pennsylvania State University
The Graduate School

DESIRABLE DIFFICULTIES IN LEARNING MULTIWORD UNITS
IN A SECOND LANGUAGE:
EXPLORING PROCESSING AND RETRIEVAL THROUGH
BEHAVIOR AND BRAIN POTENTIALS.

A Dissertation in
Spanish and Language Science

by
Manuel Pulido-Azpíroz

Submitted in Partial Fulfillment
of the Requirements
for the Degree of
Doctor of Philosophy

May 2020
The dissertation of Manuel F. Pulido was reviewed and approved by the following:

Paola E. Dussias  
Professor of Spanish, Linguistics, and Psychology  
Head, Department of Spanish, Italian, and Portuguese  
Dissertation Adviser  
Chair of Committee

Matthew T. Carlson  
Associate Professor of Spanish and Linguistics

Judith F. Kroll  
Distinguished Professor of Language Science at University of California, Irvine

Rena Torres Cacoullos  
Professor of Spanish and Linguistics

Adriana (Janet) van Hell  
Professor of Psychology and Linguistics
ABSTRACT

In this dissertation, three experiments were conducted to examine the cognitive mechanisms engaged under different conditions of retrieval practice and their association with learning outcomes. Grounded on recent proposals of desirable difficulties in vocabulary learning (Bjork & Kroll, 2015), these experiments sought to test the hypothesis that, in order to process and retrieve second language (L2) conventional multiword units efficiently, L2 speakers must learn to inhibit the L1 equivalents, which carry the same meaning, and have the same syntax but only partially overlapping lexical make-up.

Experiment 1 tested Spanish learners of English on a new paradigm that aimed to induce interference from the native language during lexical selection in a second language, as a way to train regulation of the dominant language. The results from immediate and delayed L1-to-L2 translation tests showed that recall rates were significantly higher in the group of learners that practiced in conditions of L1-interference. Faster RTs from vocal responses during practice showed more efficient lexical selection in those same learners. Additionally, RTs revealed that the more successful learners in both groups incurred a cost in accessing verb choices congruent with the native language, a finding that is consistent with an inhibitory account.

In Experiment 2, event related potentials (ERP) were used to investigate the mechanisms underlying lexical retrieval and selection in conditions of high and low conflict, similar to those elicited in the first experiment. Native and non-native English speakers were asked to select familiarized target verb–noun sequences (e.g., eat breakfast) when given two choices. Trials were either low-conflict, with only one plausible candidate (e.g., eat – shoot – breakfast) or high-conflict, with two plausible verbs (e.g., eat – skip – breakfast). While costs in response selection were modulated by plausibility and order in native speakers, only consistent effects due to plausibility were found in non-native speakers. Additionally, brain activity was only significantly
correlated with performance in native speakers. The results suggest largely similar basic mechanisms, but also that different resources and strategies are engaged by non-native speakers when resolving conflict in the weaker language, with a greater focus on individual words than on multiword units.

Taking direction from the first two studies, Experiment 3 aimed to directly examine the association between retrieval practice conditions, neurophysiological activity, and learning outcomes. English-speaking learners of Spanish completed a learning, practice and testing paradigm similar to the one used in the first experiment. The results revealed different neurophysiological signatures for participants in each of the two practice conditions. Learners who saw unrelated distractors presented a greater N400 for incongruent relative to congruent collocations, as well as a response-selection related right frontal effect (RFE). However, learners who saw L1-related distractors presented no RFE, but revealed instead a reversed N400 effect – i.e., a more negative peak for congruent than for incongruent collocations –, indicating L1 inhibition. The analyses revealed a direct association between ERPs during practice and learners’ ability to recall collocations, such that greater inhibition in congruent ERP trials predicted higher accuracy in the immediate post-tests.

Taken together, the results indicate that manipulating the degree and type of conflict elicited under different practice conditions results in the engagement of different cognitive mechanisms. The evidence from behavior and neurophysiological responses support the hypothesis that practice in regulating activation of the native language is a desirable difficulty that enhances L2 learning.
# TABLE OF CONTENTS

LIST OF FIGURES .......................................................................................................................... viii

LIST OF TABLES ............................................................................................................................. x

ACKNOWLEDGEMENTS ................................................................................................................ xi

Chapter 1  Introduction ....................................................................................................................... 1

Language non-selective lexical access in multiword units ................................................................. 4
The role of the L1 in learning L2 collocations ................................................................................. 5
Interference in L2 memory representations and the role of cognitive control ......................... 7
The role of lexical competition in retrieval and selection of L1 and L2 of multiword units .......... 8
Conflict-related ERP components ................................................................................................. 10
Language regulation and interference inhibition as desirable difficulties in L2 learning .................. 13
Roadmap to dissertation .............................................................................................................. 18

Chapter 2  Desirable difficulties while learning collocations in a second language:
Conditions that induce L1 interference improve learning ............................................................... 19

Introduction ................................................................................................................................... 19
Research questions ...................................................................................................................... 20
Predictions .................................................................................................................................... 21
Method .......................................................................................................................................... 22
Participants ................................................................................................................................. 22
Individual differences measures ............................................................................................. 22
Materials ....................................................................................................................................... 24
Experimental procedure .......................................................................................................... 26
Familiarization phase ............................................................................................................... 26
Practice ........................................................................................................................................ 29
Testing ......................................................................................................................................... 30
Results of recall tests .................................................................................................................. 32
Accuracy in recall ...................................................................................................................... 32
Reaction time analysis ................................................................................................................ 35
Discussion ...................................................................................................................................... 40
Enhanced learning of incongruent collocations ........................................................................... 42
Native language regulation as part of L2 learning .................................................................. 43
Evidence of L1 inhibition during learning .............................................................................. 45
Regulation of lexical competition within the L2 ...................................................................... 46
Types of errors ........................................................................................................................... 47
Individual differences in memory ............................................................................................ 48
The role of cognitive control in learning incongruent collocations ....................................... 49
Generalizability of efficient learning ...................................................................................... 49
Conclusion .................................................................................................................................... 50
Chapter 3  The neural correlates of conflict detection and resolution during multiword lexical selection: Evidence from bilinguals and monolinguals

Introduction .................................................................................................................................51
Materials and Methods ................................................................................................................54
  Participants .................................................................................................................................54
  Materials ..................................................................................................................................54
  Procedure ..................................................................................................................................60
  EEG recording and analysis .........................................................................................................61
Predictions ......................................................................................................................................63
  Predictions for non-native speakers ..........................................................................................64
Behavioral results ..........................................................................................................................65
  Accuracy in the Lexical Selection task .......................................................................................66
  Reaction Times in the Lexical Selection task ...............................................................................69
  Discussion of behavioral results ................................................................................................70
ERP results .....................................................................................................................................72
  ERP analysis of native English speakers ..................................................................................72
  Discussion of ERP results in native speakers .............................................................................77
  ERP analysis of Spanish-English bilinguals ...............................................................................81
  Discussion of ERP results in non-native speakers ....................................................................85
Discussion .......................................................................................................................................87
Conclusions .....................................................................................................................................92

Chapter 4  Connecting practice conditions, neurophysiology and learning: training native language inhibition improves L2 lexical retrieval

Introduction .....................................................................................................................................94
  The effect of practice modality during lexical selection on learning ............................................95
  The effect of test format on retrieval accuracy ..............................................................................97
  Exploring the generalization of benefits derived from desirable difficulties .............................97
Research Questions .......................................................................................................................98
Predictions ......................................................................................................................................98
Methodology .................................................................................................................................99
  Participants .................................................................................................................................99
  Individual differences measures ................................................................................................100
  List of collocations .....................................................................................................................102
  Experimental Procedure .............................................................................................................103
  Familiarization materials.............................................................................................................103
  Familiarization procedure ...........................................................................................................103
  Practice materials ......................................................................................................................105
  Practice procedure .....................................................................................................................106
  Testing materials .......................................................................................................................107
  Test administration, scoring and analysis ....................................................................................110
  EEG recording and analysis ........................................................................................................112
Results ..........................................................................................................................................114
  Results of ERP from Practice ....................................................................................................114
  Discussion of ERP results ..........................................................................................................121
  Results of Translation and Cloze tests .......................................................................................124
LIST OF FIGURES

Figure 1-1: Results of immediate and delayed retention tests in Schneider, Healy and Bourne (2002). ................................................................. 14

Figure 2-1: Sequencing and feedback in the Study procedure. .................................................. 28

Figure 2-2: Results of accuracy rates in immediate and delayed recall tests, for each Type of collocation ................................................................................. 34

Figure 2-3: Growth Curve Analysis of RTs for verb selection across the three practice sessions ................................................................. 37

Figure 2-4: Whole session RT averages for each of the collocation sublists ............................................. 39

Figure 2-5: Recall accuracy rates are for the one-month delayed test .................................................. 40

Figure 3-1: Sample trial of the Lexical Selection Task ............................................................................ 59

Figure 3-2: Accuracy in selection of the target verb by condition ......................................................... 68

Figure 3-3: RTs for selection of the target verb by Group ..................................................................... 70

Figure 3-4: ERP waveforms for native speakers at (a) FC6 (right frontocentral) and (b) CP1 (left centroparietal) .................................................................................. 76

Figure 3-5: Scalp topographies of the difference waves showing the effects of (a-b) Distractor Plausibility (Plausible – Implausible) and (c-d) Order (TD – DT in plausible trials) in the 500-600 and 650-800 ms time windows, respectively ........................................ 76

Figure 3-6: ERP waveforms for non-native speakers at (a) FC6 (right frontocentral) and (b) CP1 (left centroparietal) .......................................................... 84

Figure 3-7: Scalp topographies of the difference waves in non-native speakers, showing the effects of (a-b) Distractor Plausibility (Plausible – Implausible) and (c-d) Order (TD – DT in plausible trials) in the 600-800 and 800-950 ms time windows, respectively .......................................................... 85

Figure 4-1: Sequencing in the Study procedure .................................................................................. 104

Figure 4-2: Sample sequence of the Practice procedure ......................................................................... 107

Figure 4-3: Grand average ERPs of the Unrelated group .................................................................. 118

Figure 4-4: Grand average ERPs of the L1-Interference group .......................................................... 119
Figure 4-5: Scalp topographies of the difference waves showing the effects of Congruency (Incongruent – Congruent) and Order (DT – TD in plausible trials) in the 300-500 and 600-800 ms time windows for each experimental group. .......................... 120

Figure 4-6: Correlations between individual N400 mean differences at Cz and recall accuracy in immediate post-tests.......................................................... 120

Figure 4-7: Results of recall accuracy in immediate and delayed post-tests across the two experimental groups ............................................................... 126
LIST OF TABLES

Table 2-1: Summary of cognitive and proficiency measures for each group. .......................... 23
Table 2-2: Sample of collocations matched across languages and sublists. ......................... 25
Table 2-3: Testing protocol. ........................................................................................................ 28
Table 3-1: Sample of experimental manipulation (eat breakfast). ......................................... 53
Table 3-2: List of Verb-Noun phrases per condition. ............................................................... 56
Table 3-3: Summary of cognitive and proficiency measures....................................................... 66
Table 3-4: Results of accuracy in selection of the target verb by condition. ......................... 67
Table 3-5: Reaction Times in selection of the target verb by condition. ............................... 69
Table 4-1: Summary of cognitive and proficiency measures................................................... 101
Table 4-2: Sample of practice collocations matched across languages and sublists............. 105
Table 4-3: Generalized linear mixed effects model output ....................................................... 126
ACKNOWLEDGEMENTS

This dissertation was only possible thanks to the involvement of a large group of people: from colleagues and students who, during years of teaching languages, inspired me to pursue some of the questions explored here; to the world-class faculty and always-willing-to-help colleagues and mentors at Penn State; to the learners and bilingual speakers who invested their time in participating in research; to family and friends whose support never fails. Every part of this work is in some way indebted to one of you. It really does “take a village” – and a global one.

I owe many thanks to my dissertation committee: Giuli Dussias, Judy Kroll, Matt Carlson, Rena Torres Cacoullos and Janet van Hell. I have been exceptionally fortunate to work with you. A large number of the ideas in this dissertation emerged, and were then developed, through thought-provoking discussions with you in the classroom and beyond. Since much of the ensuing work that took place in research laboratories, I am indebted to Judy Kroll, who generously allowed me to first join her lab and gain the necessary training to conduct research using EEG. I would like to extend my utmost gratitude to Giuli Dussias, my doctoral adviser, for her unwavering dedication and guidance over this time. She has been a model of brilliance, energy, empathy and love for research. I hope her influence will continue to inspire me for years to come.

I would also like to thank my graduate school colleagues and friends, both in Spanish linguistics and literature. I am grateful to many friends and colleagues in the Center of Language Science with whom I have had the pleasure to collaborate, and who have in countless occasions provided help, advice and support: Annie Beatty-Martínez, Grant Berry, Katrina Connell, Isabel Deibel, Johan De La Rosa, Alba García Alonso, Mary Henderson, Mike Johns, Rafał Jończyk, Hossein Karimi, Priscila López Beltrán, Alex McAllister, Rhonda Mudry, Christian Navarro Torres, Laura Rodrigo Cristóbal, Patricia Román Fernández, Jonathan Steuck, Andrea Takahesu Tabori, Giulia Togato and Haoyun Zhang. My graduate experience would not have been nearly as enjoyable without friends who have made central Pennsylvania feel more like home throughout these years: Ricardo Andrade, Antonella Fonseca, David Dopfél, Marylin and Don Keat, Daniel Ramirez-Smith and Alexis Santos-Losada.

A very special thanks goes to my family (los Pulido-Azpíroz y los Pérez de Zabalza-Vidaurre), whose unrelenting support and confidence has always accompanied me. Above all, my gratitude goes to my wife, Lucía, who is there to celebrate every accomplishment as well as to overcome every obstacle, no matter how seemingly insurmountable. Thanks to you, this and every journey is a more meaningful and joyful one.

Finally, I would also like to thank funding agencies that have generously supported this work. These include Penn State’s College Liberal Arts, which supported me through a release to write my dissertation as well as research grants, including a RGSO Dissertation Support award; a National Science Foundation Doctoral Dissertation Improvement Award to Paola E. Dussias and Manuel F. Pulido (NSF-BCS-1844188); a Doctoral Dissertation Grant journal awarded to Manuel F. Pulido by the journal Language Learning; and grant NSF-OISE-1545900 awarded to Paola E. Dussias.¹

¹ The findings and conclusions in this dissertation do not necessarily reflect the view of the funding agencies.
Chapter 1

Introduction

Evidence from neurophysiological studies has shown that second language (L2) users, may in some cases show neural signatures similar to those of native speakers (Morgan-Short et al., 2012; Steinhauer et al., 2009) even when acquiring a foreign language after adolescence. However, despite the fact that some adults may attain native-like proficiency in an L2, success in learning a second language is still largely variable. Such variability is particularly evident in aspects of the L2 that are only partially congruent across the two languages (for a review, see Tolentino & Tokowicz, 2011). In fact, brain event related potentials (ERPs) have evidenced that learners show sensitivity to aspects of the L2 that are either entirely congruent to the L1 or entirely unique to the L2, but not to aspects of the second language which partly differ across the two languages (Foucart & Franck-Mestre, 2011; Osterhout et al., 2006; Sabourin & Haverkort, 2003; Tokowicz & MacWhinney, 2005). Partial L1-L2 incongruency appears to be particularly problematic for L2 learners at various levels, whether morphosyntactic (Foucart & Franck-Mestre, 2011; Sabourin & Haverkort, 2003; Tokowicz & MacWhinney, 2005), phonological (Barrios, Jiang, & Idsardi, 2016; Brown, 2000) or lexical (Peters, 2016; van Hell & Tanner, 2012).

An aspect of language that shows large variability in learning success due to partial L1-L2 congruency is that of collocations, defined as combinations of words that are strongly associated because of their tendency to appear together (Gries, 2013). This dissertation investigates the cognitive mechanisms that must be engaged for successful learning, and the learning conditions that enable the learning of L1-L2 incongruent collocations, i.e., those that are only partly congruent with their counterparts in the native language. For instance, a verb-noun phrase (V-NP) sequence like run a store is collocational in English, while the Spanish translation equivalent carry a store is not. According to recent proposals (e.g., Arnon & Snider, 2010; Durrant & Doherty, 2010; Siyanova-Chanturia et al., 2017), the mental
representation of collocations is different from that of single lexical items, but also from other fixed multiword expressions, such as idioms (e.g., *kick the bucket*). Further, collocations have abstract internal syntax, and are said to influence language acquisition and language processing (Ibbotson, 2013).

Knowledge of L2 collocations is a critical asset for non-native speakers trying to master a foreign language. Learning them helps L2 speakers fulfill pragmatic functions and produce output that matches the expectations of native interlocutors (Bardovi-Harlig, 2009). Moreover, collocations are important because they facilitate processing in both native speakers (Durrant & Doherty, 2010) and second language speakers (Wolter & Gyllstad, 2011, 2013), and also facilitate the production of utterances that meet the linguistic expectations of interlocutors (Boers, Eyckmans, Kappel, Stengers & Demecheleer, 2006; Saito, 2019). Although much is known about the special status that conventional speech has in the lexicon of native speakers, the cognitive mechanisms underlying how collocations come to be represented in adult L2 learners are not well understood. We know that L2 speakers are able to learn new single vocabulary items throughout life; they are, however, less successful at learning and using conventionalized collocations, even when collocations are composed of lexical items that, individually, are well known (Nguyen & Webb, 2017).

Despite the documented constraints associated with learning collocations in an L2 (e.g., Nesselhauf, 2003, 2005; Nguyen & Webb, 2017; Peters, 2016), some learners still manage to successfully achieve native-like performance in processing L2 collocations, but the evidence on what it takes for a learner to be successful, or for a learning context to be enabling, is mixed. Past work has shown that congruent collocations (i.e., those with the same lexical items in the L1 and L2) do not necessarily pose serious learning and processing challenges for L2 speakers; however, the development of L1-L2 incongruent representations is highly problematic (Nguyen & Webb, 2017; Peters, 2016). Even when learned, incongruent collocations still give rise to processing costs (Wolter & Gyllstad, 2011, 2013). Studies in which incongruent collocations have been specifically targeted in instruction (Boers et al., 2017; Peters, 2012, 2016) have produced limited gains. Some studies suggest that
Collocations can be learned incidentally through repeated exposure (Pellicer-Sánchez, 2017; Webb, Newton & Chang, 2013), but enough incidental exposure may not always be possible in contexts of classroom-based learning with limited L2 input. A combination of input flood and text enhancement (e.g., underlining the target collocations) has been shown to lead to greater success in form recognition, but less so much in the ability to recall and produce collocations (Szudarski & Carter, 2016). Prior studies have also suggested that input conditions that cause interference should be avoided (Boers et al., 2014, p. 65). While this seems like a sensible approach, evidence also indicates that learners experience cross-linguistic interference even if it is not overtly present in the input (Yamashita & Jiang, 2010). Having a high level of cognitive resources or being immersed in the L2 appear to contribute to positive outcomes, but not uniquely (Segalowitz & Freed, 2004; Sunderman & Kroll, 2009).

The work in this dissertation takes direction from a recent proposal in the learning literature (Kornell, Hays & Bjork, 2009; Bjork & Kroll, 2015)—that creating “desirable difficulties” during early stages of learning can result in greater gains—to test a new hypothesis about the processing of L2 collocations, one that stems from the observation that classroom-based studies have likely provided appropriate input, but have generally failed to engage the critical cognitive mechanisms that support the learning and retrieval of incongruent collocations. The hypothesis is that to process incongruent L2 collocations efficiently, L2 speakers must learn to inhibit the equivalent L1 collocations, which carry the same meaning, and have the same syntax but only partially overlapping lexical make-up (see Bogulski, Bice & Kroll, 2018 for a similar recent hypothesis about single word learning). These features—syntactic and lexical overlap—have been hypothesized in related L2 research to increase co-activation of L1 collocational competitors. By training lexical selection in conditions that require L1 regulation, learners may be able to develop control mechanisms that are inherent to bilinguals’ daily experience of selecting between candidates in two languages. The approach employed here capitalizes on previous findings from the memory and cognitive control literature to investigate the representation of incongruent collocations.
from a novel perspective. Specifically, rather than assuming facilitation due to cross-language congruency (Wolter & Gyllstad, 2011), the goal is to examine the prevalence of L1 interference and its effect on lexical selection.

Language non-selective lexical access in multiword units

While multiword units are by definition idiomatic and express a holistic meaning, they are also often analyzable into their individual constituents (Bybee, 2010, pp. 25-28; Langacker, 1987). The literature on idioms has provided evidence that the meanings of individual words are also accessed during the processing of idioms (e.g., Gibbs, 1980; Hamblin & Gibbs, 1999; Sprenger, Levelt & Kempen, 2006; Titone & Connine, 1999). Findings show that greater decomposability (i.e., the extent to which the meaning of a multiword unit can be decomposed based on its individual words) facilitates integration in context when an interpretation must be selected between competing literal or idiomatic meanings (Libben & Titone, 2008; Titone & Libben, 2014). At the same time, there is ample evidence that when idioms can be interpreted literally, there is interference with the non-literal meaning (Caillies & Declercq, 2011; Colombo, 1993). Therefore, while there is evidence of the fine-grained storage of multiword units in the mind, findings demonstrate that their mental representation often relies on syntactic parsing and on more than one level of semantic representation. Furthermore, the extent to which multiword units are decomposable and transparent is complicated by the fact that native and non-native speakers may have different perceptions. Specifically, the cultural and linguistic background of learners is likely to affect their perceived aspects such as the transparency of multiword units (Boers & Webb, 2015).

The fact that multiword units are analyzable has important implications for L2 learners. It is not only relevant to ask whether conventional expressions in the L1 have equivalent expressions in the L2, but also whether the lexical make-up of such expressions is congruent across languages. In an L2, accessing collocations that are equivalent (i.e., lexically
congruent) with the L1 appears to be unproblematic. But recent research has shown that
colloctions that are incongruent across the L1 and L2 present difficulties in processing, even
when they have been learned (Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2015).
While the representation of L2-specific multiword units has been acknowledged to be
problematic for decades (e.g., Irujo, 1984; Nguyen & Webb, 2017; Peters, 2016), the role of
previous L1 experience in forming these representations is still not well understood. Wolter &
Gyllstad (2011) conducted one of the first studies (see also Yamashita & Jiang, 2010) to
systematically investigate the effect of congruency on the processing of L2 collocations.
Their results showed that not only were Verb-Noun (V-NP) collocations (e.g., pay a visit)
processed faster than unrelated V-NP combinations (e.g., do a visit), but also that collocations
that were congruent across the speaker’s two languages were facilitated.

The results were interpreted as an effect of facilitation in processing congruent
relative to incongruent collocations. Facilitation accounts assume either “doubled activation”
of lexical-semantic nodes across the L1 and L2, or an age-of-acquisition effect due to earlier
acquisition of congruent collocations (Wolter & Gyllstad, 2011; Wolter & Yamashita, 2015).
An alternative, non-exclusive hypothesis explored here, is that congruent collocations do not
necessarily facilitate online processing; instead, we propose that the implicit activation of L1
lexical links causes interference when L2 speakers process and retrieve incongruent
colloctions.

The role of the L1 in learning L2 collocations

Ample research demonstrates that the degree of similarity between the L1 and the L2
influences the ability to learn and process both the lexicon of a second language (e.g.,
Brenders, van Hell & Dijkstra, 2011; Dijkstra, Miwa, Brummelhuis, Sappelli & Baayen,
2010; Dijkstra, van Hell & Brenders, 2015) as well as its structure (e.g., Morett &
MacWhinney, 2013; Sasaki, 1991; Tolentino & Tokowicz, 2011). Importantly, the presence
or absence of L2 structural features in learners’ L1 also impacts the effectiveness of the type
of instruction provided. What the extant research shows is that L2 features that are absent from learners’ native languages require special attention and are best learned when differences are highlighted (Tolentino & Tokowicz, 2014). A relevant perspective in accounting for differences in learning L2-specific structures is the notion of blocking in associative learning (Kamin, 1968), and the idea that entrenched knowledge in the native language may block the formation of L2 representations (for applications of blocking in L2 learning see, e.g. Ellis et al., 2012; Ellis & Sagarra, 2010).

In what concerns the acquisition of L2 collocations, there is ample evidence that L2 learners often rely on L1 knowledge of how word meanings are combined to learn collocations in the L2 (Bahns, 1993; Biskup, 1992; Irujo, 1984; Granger, 1998; Nesselhauf, 2003, 2005). The discussion in the previous section shows that, relative to congruent collocations, incongruent collocations are at disadvantage in learning and processing. Data from studies on L2 production provide further insight into the role of the L1. A revealing piece of evidence comes from the finding that literal translations (calques) in learners’ L2 output that are not idiomatic are often derived from the learners’ L1. In a study that examined the use of collocations in the free written production of advanced German learners of English, Nesselhauf (2003) described that L1-based translation could explain numerous non-idiomatic choices for the verbs, nouns and prepositions in collocations and idioms. She argued that the evidence of L1 influence, even at higher levels of proficiency, suggests that it is not enough to teach target L2 multiword units, and that L1 and L2 forms should be explicitly contrasted (Nesselhauf, 2003, p. 239). A classroom-based study by Laufer and Girsai (2008) tested the effectiveness of the L1-L2 contrastive approach in Hebrew-speaking high-school learners of English. In their study, three groups of learners were initially exposed to the same text and responded to comprehension questions. One day later, one group completed meaning-focused questions that involved using the target collocations; a second group received non-contrastive form-focused questions, including multiple-choice and fill-in exercises; the third group completed questions that involved contrasting and translating from the L1 into the L2 and vice versa. The results of immediate and one-week delayed post-tests showed that the
translation treatment in the third group produced the highest rates of learning. Moreover, the rates of learning achieved by the learners in the contrastive treatment seemed to outperform other studies targeting the learning of collocations (although, as the authors acknowledge, comparisons are not always straightforward; Laufer & Girsai, 2008, p. 710). The results from the study suggest that learning can be enhanced through approaches based on contrasting the L1 and L2. However, the results of Laufer and Girsai (2008) provide no insight into the specific mechanisms of retrieval, beyond the fact that an explicit association was established between existing L1 and novel L2 representations. More importantly, it is unclear whether such explicit awareness is directly responsible for the improved performance in the L2 and may facilitate retrieval that is unmediated by contrastive associations with the L1. The focus of the present research is on investigating the development of the monitoring and selection mechanisms believed to allow for successful and more efficient retrieval and recall.

**Interference in L2 memory representations and the role of cognitive control**

Findings from research on phonological, syntactic and semantic priming studies have provided convincing evidence that access to linguistic representations at different levels is language non-selective (e.g., Carroll & Conklin, 2017; Hartsuiker, Beerts, Loncke, Desmet & Bernolet, 2016; Thierry & Wu, 2007). A consequence is that both languages become activated in parallel, and so L2 speakers must learn to select among competing alternatives available in both their languages. The need to regulate competition is believed to place increased demands on cognitive control in bilinguals (Bialystok, Craik, Green & Gollan, 2009). Moreover, the evidence suggests that resolving competition across languages comes at a cost (e.g., Hoshino & Thierry, 2011). When selecting among competitors, the controlled retrieval of a target candidate renders the alternative not selected more difficult to access (Anderson, Bjork & Bjork, 1994). What research on bilingualism has shown is that L2 speakers develop enhanced ability to monitor between conflicting representations (e.g., Abutalebi, 2008), necessary for controlled rather than for automatic retrieval. We suggest that
learners that succeed in learning L1-L2 incongruent constructions are those that can engage control and regulate competition from the L1 (although this discussion concerns lexical learning, for similar arguments regarding general language ability in children see White, Alexander & Greenfield, 2017; L2 proficiency development, Linck, Osthus, Koeth & Bunting, 2014). The proposal in this dissertation is also informed by the work on paradigms requiring controlled selection, such as the Stroop effect (e.g., name red in the visually-presented word “blue” printed in red), where participants must inhibit related interfering representations in order to selectively control a less accessible response, to train the ability to select the appropriate representations in L2 learners. The association between previous conflict-related studies (particularly in the Stroop paradigm) and the study of conflict in lexical selection is discussed in the following section.

The role of lexical competition in retrieval and selection of L1 and L2 of multiword units

The importance of making acceptable lexical choices is an aspect that non-native speakers become acutely aware of in any occasion in which knowledge from the L1 is found to not be directly applicable to the other language. For instance, while an individual can be said to skip breakfast in English, its idiomatic Spanish equivalent saltarse el desayuno “jump the breakfast”, if literally translated, would present a native English speaker with a very unusual phrase. A growing number of studies supports the view that conventional word choices like skip breakfast are not retrieved as independent items, i.e., skip + breakfast, but rather as pre-stored connections between words, or multiword units (Arnon & Snider, 2010; Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2015). Multiword units offer valuable insight into the role of automatic lexical activation as based on the strength of experience-based lexical associations. Familiar multiword units come online when a given target is accessed, i.e., breakfast, which then may activate common collocates, e.g., eat but also non-target candidates such as skip, serve, or prepare.
In order to investigate conflict detection that results from interference between lexical candidates, one must first understand when interference during selection of an item in a given context may arise. However, research on the specific mechanisms involved in detecting and resolving conflict in actual language usage is surprisingly lacking. While some paradigms such as Stroop have been used to investigate lexical selection at the individual word level, there is a notorious absence of work investigating how lexical selection occurs in units greater than a single word (with the exception of studies investigating selection of the determiner given a noun, e.g., Dhooge, De Baene & Hartsuiker, 2016; Schriefers, 1993; Schriefers, Jescheniak & Hantsch, 2002; Schriefers, Jescheniak & Hantsch, 2005). The classic Stroop task is an excellent paradigm to explore the cognitive mechanisms involved in conflict detection and resolution during lexical selection, given its simplicity in design. Moreover, the mechanisms elicited by the task are assumed to be domain general mechanisms, common to other cognitive tasks (see Larson, Clayson, & Clawson, 2014 for a review of ERP findings on Flanker and Stroop), and directly related to selection in language use (Jackendoff, 1995; Siyanova-Chanturia, Conklin & van Heuven, 2011; Siyanova-Chanturia, Conklin & van Heuven, 2017). But while the simplicity of the Stroop task has provided a valuable tool into the mechanisms of conflict detection, it is obviously not without its limitations. Perhaps most preeminent is the fact that word selection in actual linguistic performance does not occur in the form of independent lexical choices; rather, choices are made to connect words with other words to form phrases that themselves are then nested into sentences, and these in turn into broader discourse. Critically, words to be selected must be an acceptable lexical choice based on the surrounding context. Put simply, only words that “fit” the syntactic and semantic context need to be selected.

An exception to the general absence of research focusing on lexical selection in multiword units is made by a number of studies examining selection of the determiner. These studies have usually employed the picture-word interference paradigm (Schriefers, Jescheniak & Hantsch, 2002; Schriefers, Jescheniak & Hantsch, 2005; West, 2003; West & Alain, 1999). The common finding is that there are slower naming latencies when the determiner of the
distractor noun is incongruent with the determiner of the target, consistent with a cost due to conflict during selection. Dhooge et al. (2016) recently used ERPs to investigate the timing of determiner selection. In their experiment, participants were required to name a picture or withhold their response based on the determiner for the noun, using a paradigm that capitalized on the go-nogo N200 component. The results suggested that distractor nouns with an incongruent gender interfered during selection of the target noun. But because the manipulation was designed to elicit differences in the go-nogo N200, which reflects response inhibition, no conclusions about conflict detection during selection can be made.

Neurophysiological measures are ideal to investigate this question because they will allow for the examination of the ERP components associated with selection mechanisms and their time-course in both native and non-native speakers. Monolingual speakers would be thought to experience interference stemming from within-language spreading activation that needs to be resolved. However, this problem is exacerbated in bilinguals, who when using another language also experience cross-language activation (Thierry & Wu, 2007; Oppenheim, Wu & Thierry, 2018; Wu & Thierry, 2010). Because lexical selection involves detecting conflict between candidates concurrently available for selection, understanding the process of lexical selection requires investigating the conflict detection and resolution mechanisms involved.

**Conflict-related ERP components**

In order to examine the issue at hand, we review the literature on related paradigms that have explored ERP components associated with conflict detection. Since the question of conflict detection has not been explored in connection with lexical selection in multiword units, we turn to the literature on conflict detection in widely researched cognitive tasks. As this body of research demonstrates, while many tasks rely on similar basic domain-general mechanisms, the particulars of each paradigm critically affect the timing of conflict detection.
Here we briefly review this literature, which evidences similarities and differences across experimental paradigms.

Several studies have identified the electrophysiological correlates of conflict detection using both linguistic and non-linguistic tasks (e.g., Larson, Clayson & Clawson, 2014; Dyer, 1971; Glaser & Glaser, 1982; van Heuven & Dijkstra, 2010). While tasks such as Flanker and Stroop differ in important ways, they are also believed to rely on the same basic conflict detection and resolution mechanisms. Flanker is a non-linguistic task, in which participants must indicate the direction of an arrow, typically by pressing a button (and ignore distractor arrows in incongruent trials). Stroop is a psycholinguistic task in which participants must retrieve the name of a color and ignore other related information. Typically, participants have to name the font color in congruent (e.g., the word BLUE printed in blue font) and incongruent trials (e.g., the word BLUE printed in red). While both tasks may rely on the same domain-general mechanisms, there are important differences across both paradigms that affect the timing of components in ERP studies. For instance, conflict detection in the Flanker task is indexed by the N200 component, a negative deflection peaking at around 200 ms after stimulus presentation in incongruent trials. In the Stroop paradigm, on the other hand, conflict detection appears later and is indexed by the N_{inc} component, peaking at around 450 ms post-stimulus presentation over the left centro-parietal scalp, and is sometimes referred to as the N450 (Glaser & Glaser, 1982; van Heuven & Dijkstra, 2010; Coderre, van Heuven & Conklin, 2013). Although the scalp distribution of the N_{inc} is typically over left central-parietal electrodes, source localization techniques suggest that it is generated in the pre-frontal cortex by the ACC (Hanslmayr, Pastötter, Bäuml, Gruber, Wimber & Klimesch, 2008; West, 2003), which is associated with executive control. The N_{inc} is elicited after the word and color are presented at the same time in the incongruent condition relative to congruent and control trials. Given that Stroop is a linguistic task, involving semantic processing of the word and color, and that lexical access is indexed by the N400, it is only expected that conflict detection in this paradigm will be at least as slow as semantics.
A Late Positive Component (LPC) or slow positivity component, is also commonly reported in ERP studies using the Stroop paradigm. This is a positive-going wave appearing between 600 and 900 ms post-stimulus over centro-parietal regions, inverting polarity over the anterior scalp concomitantly in time. It has been suggested to be an index of response selection, on the grounds that it correlates with RT and accuracy (West, Jakubek, Wymbs, Perry & Moore, 2005). Others, however, have proposed that the LPC reflects semantic re-activation of the word (Liotti, Woldorff & Mayberg, 2000). Therefore, while it appears to be associated with the generation of a response, its functions are less clear.

One important question concerns whether the N_{inc} reflects conflict detection or conflict resolution. An insightful approach to address this question is offered by experiments that have varied the stimulus onset asynchrony (SOA) by presenting color and word stimuli at different times, to investigate the precise timing of color and word interference (Coderre, Conklin & Van Heuven, 2011; Dyer, 1971; Glaser & Glaser, 1982). Manipulating the SOA of the word relative to the color stimulus allows for the dissociation of automatic conflict detection from the need to resolve conflict to produce a response. In a recent ERP study, Coderre et al. (2011) used long SOA manipulations of -400 ms, 0 and +400 ms. Because RTs in manual Stroop studies have been shown to occur at around 500 ms, a response is already being prepared or made by the time the word is presented in the +400 SOA condition. The authors predicted that if the N_{inc} is an index of conflict resolution, is should be absent in the positive SOA after a response has been made; if it indicates conflict detection, however, it should still be elicited. The results of the study found that the N_{inc} was still present at 700-850 ms after presentation of the color and after a response was made. This was the case even in fast-response trials, in which responses had been recorded by 600 ms. The same type of SOA manipulation has provided further insight into the LPC, showing that its presence is determined by the need to execute a response. In an EEG Stroop study with short SOA manipulations, Appelbaum et al. (2009) found that in the -100 ms and -200 ms conditions, the LPC appeared earlier by the same amounts of time, respectively, and was similarly delayed in the +100 SOA condition. The longer SOA used in Coderre et al. (2011) revealed a shift
window in the -400 ms condition, but no LPC when presentation of the word was delayed by 400 ms.

The question of how conflict detection in the Stroop task may be affected by bilingualism has been addressed in some prior research. Several ERP studies give compelling support to the idea that L2 lexical access is slower relative to the L1 (for an overview see van Heuven & Dijkstra, 2010). More specific to the Stroop task, a previous study found that Chinese-English bilinguals were slower by a difference of 100 ms in their L2 (Coderre et al., 2013). The authors investigated two aspects of conflict detection hypothesized to have opposite effects in bilinguals. On the one hand, bilingual speakers should experience slower lexical access in their L2; on the other, based on research suggesting a bilingual advantage in cognitive control, bilinguals should be faster in detecting conflict. Overall, the results showed a delay when the task is performed in the L2 and presented some evidence for a bilingual advantage modulated by L2 proficiency. Interestingly, the results from the Chinese-English speakers did not give support to the idea of a bilingual L2 lexical disadvantage hypothesis, by which lower L2 proficiency should result in less interference due to weaker lexical access. Rather, lower proficiency resulted in a larger interference effect in the L2 English. The authors attributed these differences to a bilingual cognitive advantage, with a lesser cost in the higher-proficiency group. Therefore, the evidence to date highlights the variety of factors specific to bilinguals that may impact performance in the Stroop paradigm.

**Language regulation and interference inhibition as desirable difficulties in L2 learning**

Based on the above discussion, the ability to resist interference from the native language would be thought to play a nontrivial role in L2 users’ ability to learn L1-L2 incongruent multiword units from the input, and subsequently retrieve those representations. In this sense, the ability to handle competition from the L1 can be regarded as a desirable skill for learners to develop. This idea is grounded, first, on abundant evidence that some bilinguals show superior performance in indices of cognitive control, often measured through
tasks that require the suppression of interfering information (e.g., see Bialystok, Craik, Green, & Gollan, 2009 and Costa, Hernández, Costa-Faidella, & Sebastián-Gallés, 2008 for reviews; Bialystok, Kroll, Green, MacWhinney, & Craik, 2015, discuss the bilingual advantage related controversy). Additionally, studies on language learning have suggested that conditions that require training in suppression competition from the L1 result in greater gains (Bogulski, Bice & Kroll, 2018; Schneider, Healy & Bourne, 2002). Previous studies in L2 vocabulary learning give support to this idea. In a study in which native English speakers were tasked with learning French words, Schneider, Healy and Bourne (2002) manipulated the direction in which participants studied the words. While some participants studied the translation pairs from English to French, others saw French to English. When tested at the end of the session, the results showed that the French-to-English condition (L2-to-L1) performed better (see Fig. I-1). This pattern was not surprising, given the fact that encoding direction and retrieval should be easier when required to respond in the L1, as opposed to having to produce a new word in the other language. After a one-week delay, however, participants were tested again. The results showed that despite the initial difficulty in the group who had to map and produce words from the L1 to the L2, the pattern was reversed, with learners in the more difficult condition outperforming the other group. The finding that the more challenging L1-to-L2 direction resulted in improved long-term retention is consistent with the idea that remembering in the weaker language was costly at first, but that the ability to retrieve L2 words was improved with practice.

Figure I-1: Results of immediate and delayed retention tests in Schneider, Healy and Bourne (2002).
In a more recent study, Bogulski, Bice and Kroll (2018) tested the hypothesis that language learners develop the ability to regulate competition from the native language. Bogulski and colleagues proposed that if bilinguals are able to better suppress interference from the native language, benefits in retrieval might be observed. To test this idea, they included four groups of participants who studied Dutch words: a group of English monolinguals, as well as L1 English–L2 Spanish, L1 Spanish–L2 English and L1 Chinese–L2 English bilinguals. Because establishing direct mappings through the L1 might be enough to improve learning, the group of native English speakers provided a baseline. The results showed that only English-Spanish bilinguals outperformed the English monolingual group. That is, bilinguals whose first language was not English did not differ from the monolingual group, but English-Spanish bilinguals, who also learned through their L1, outperformed the English monolingual participants in a lexical decision task containing the words learned. The authors attributed the advantage to the fact that the English-Spanish bilinguals were able to better regulate competition from the L1, and proposed the L1 Regulation Hypothesis. While, as the authors acknowledged, the matching between the linguistic profiles of the groups was not perfect (p. 14), the hypothesis provides a provocative testing ground for desirable difficulties in L2 vocabulary learning.

The approach in the studies described, which sought to engage relevant cognitive mechanisms to enhance gains in learning and retention, is part of a broader approach that explores desirable difficulties (Bjork, 1994, 2018; Bjork & Kroll, 2015). The examples reviewed underscore the need to take into consideration the theoretical distinction between knowledge and performance or, put somewhat differently, between storage and retrieval (Bjork & Bjork, 1992). As illustrated by the mismatch between immediate and delayed tests, conditions that optimize storage strength (the creation of a mental image) may differ from the optimal conditions for retrieval strength (the ability to access those representations).

The conditions that have been found to improve learning more broadly also include at least four general types of manipulations. These include varying the contexts of learning, instead of keeping them constant and predictable; interleaving input on different topics,
categories or examples, as opposed to presenting the more similar elements together ("blocking"); spacing study of the same content, rather than massing; and creating opportunities to test, which often require generating responses, instead of more passively taking in the input. It is important to note that all these conditions have in common that they are thought of as engaging the relevant cognitive mechanisms that support learning. That is, while most difficulties are in themselves undesirable and counterproductive, the ones identified and described above are deemed desirable in spite of the fact that they often seem to impose an initial burden.

These effects have not only been shown to produce robust and replicable effects across numerous studies but, importantly to the topic at hand, have in some cases been successfully applied to the learning of the lexicon. They should therefore be considered as desirable traits to be incorporated in the design of studies targeting L2 learning when relevant. But rather than simply replicating the exact conditions previously found, successful experiments have sought to create conditions that engage the same or similar mechanisms thought to improve learning. For example, the rationale for the testing effect is that the mere act of retrieval is a powerful “memory modifier” (Bjork, 1975). Laboratory studies have demonstrated that the need to retrieve a response during testing, or even simply attempting to retrieve the information, aids learning to a greater extent than repeated exposure to the same content (Carrier & Pashler, 1992). Because retrieving a word usually implies producing it, in language studies this is often referred to as the production effect (Forrin & MacLeod, 2017; Ozubko & MacLeod, 2010; Zormpa, Brehm, Hoedemaker, & Meyer, 2018). More specifically, producing words may improve subsequent retrieval due to several reasons. On the one hand, overt responses feed into the auditory loop. On the other hand, overt speech engages the production system, encompassing several levels of retrieval, including gesture and motor planning. The evidence suggests that production may play an important role in learning, by enforcing retrieval of form-meaning associations at multiple levels of representation, with enduring effects even after several days (Ozubko, Hourihan, & MacLeod,
Additionally, production has been found to have a greater impact on memory when produced by oneself than when listening to another person (Forrin & MacLeod, 2017).

In a recent study, Potts and Shanks (2014) compared native English speakers’ ability to learn words in Basque across three conditions, consisting of simply reading, generating responses or choosing from among four options provided. The results showed that generating responses produced the best results, followed by choosing, and the lowest gains were found for the reading condition. The finding of the superiority of retrieval over multiple-choice replicates previous research (e.g., Carpenter & DeLosh, 2006; Glover, 1989). Nonetheless, a place for multiple-choice within the desirable difficulties approach has been more recently reclaimed. The idea is that manipulating the competitiveness of the non-target alternatives should also provide a benefit in allowing test-takers to contrast and reject incorrect information. Interestingly, practice multiple-choice quizzes have been found to improve performance not only for repeat, but also for related questions (Little et al., 2012; Little & Bjork, 2015, 2016; Bjork, Little & Storm, 2014).

A connection may also be established between research on interleaving and some studies on second language learning. One hypothesis is that interleaving forces to “reload” memories. Having retrieved A and then B, reloading A again requires not using the same retrieval route. This may account at least in part for the fact that interleaving improves long-term retention and transfer. An alternative hypothesis is that, through interleaving, one is forced to resolve interference between different items as well as to notice the similarities and differences among them. This notion is certainly familiar to studies on L2 learning which involve some sort of contrast or comparison between two languages. Relevant to the topic of this dissertation, a contrastive approach for the learning of multiword units has been not only encouraged (Nesselhauf, 2003) but also shown to improve learning in at least on classroom-based intervention (Laufer & Girsai, 2008).

The studies reviewed suggest a promising research avenue and make a compelling case for the potential of applying this approach to research on language learning, and particularly to more challenging aspects of L2 acquisition.
Roadmap to dissertation

The remaining chapters are structured as follows. Chapter 2 presents a behavioral experiment, conducted in Spain with L1 Spanish learners of English, which aimed to improve learning of L1-L2 incongruent collocations by manipulating practice conditions in two groups of participants. The critical manipulation affected the distractors presented in the L2, which were either native language equivalents or unrelated. Chapter 3 reports on an ERP study that investigated the brain components elicited by during a lexical selection task. This second experiment tested a novel paradigm that was inspired on conflict-inducing experiments such as Stroop, although adapted to selection of multiword units. Chapter 4 presents a third experiment which builds on the findings of the experiments reported in the previous two chapters. It employs a manipulation akin to the one in the first experiment, while incorporating ERPs to directly investigate the effect that manipulating practice conditions induces on brain potentials, and their association with learning as measured in behavioral tests. Each of these chapters presents the methodology, procedure, results and discussion. The fifth and last chapter is dedicated to a general discussion of the results and their implications for future studies and instruction targeting the learning of L2 multiword units, and proposes ways to translate the findings to the design of pedagogical materials for both classroom- and technology-based instruction.
Chapter 2

Desirable difficulties while learning collocations in a second language:
Conditions that induce L1 interference improve learning

Introduction

This first experiment aims to investigate the association between retrieval conditions and recall of L1-L2 incongruent collocations in an L2. Based on the literature reviewed above, we hypothesize that recall of incongruent collocations is mostly impaired by interference from the native language. As previous research on memory and recall suggests, successfully encoded representations may fail to be retrieved because related information blocks recall (Anderson, Bjork & Bjork, 1994; Bjork & Bjork, 1992). Specifically, we propose that difficulty in recall is due to more strongly active competitors (L1 collocations) blocking retrieval of weakly represented L2 collocations. Therefore, it is expected that, in order to resolve cross-language competition during retrieval, the L1 analogues of target L2 collocations may need to be suppressed. One likely explanation for the limited learning gains reported in classroom-based studies is that learners are not engaging the control mechanisms required to regulate cross-linguistic competition.

A wealth of research on cognitive control indicates that control mechanisms become engaged in conditions that require selection between competing alternatives (e.g., Botvinick, Braver, Barch, Carter & Cohen, 2001; Eriksen & Eriksen, 1974; Stroop, 1935). In lab-based experiments, contexts in which targets are presented along distractors allow for the creation of conditions that require participants to select between competitors. Although neutral or unrelated distractors generate low-level conflict, distractors that represent a plausible choice interfere with selection of the target response. We will capitalize on this effect by presenting

\footnote{Part of the content of this chapter has been published in Pulido, Manuel F. & Paola E. Dussias, (2019). Desirable difficulties while learning collocations in a second language: Conditions that induce L1 interference improve learning. Bilingualism: Language & Cognition. DOI: 10.1017/s1366728919000622. Reproduced with permission.}
each of two groups of learners with different practice conditions: one group of learners will see distractors that make plausible verbs based on L1 associations (“L1-Interference group”), while a second group will be presented with unrelated and implausible distractors (“Unrelated group”). As discussed above, collocations are believed to remain analyzable and therefore learners activate the meanings of their individual words. Conflict is hypothesized to emerge through implicit translation in the “L1-Interference” group, in which the distractor verbs will be English translations of the L1 (Spanish) equivalent. Therefore, we will use a between-subject Group manipulation (“L1-Interference” vs “Unrelated” distractors) to examine learning of three different Types of collocation (congruent, incongruent and semantically related), through performance on two immediate and three delayed recall tests. Our analysis will also consider the contribution of cognitive control by including individual scores derived from the AX-CPT and Flanker tasks, as well as measures of memory from Nonword repetition (PSTM), and Reading Span (WM).

**Research questions**

In the current study we will aim to address the following questions:

1- How does rejection of L2 distractors that are congruent with L1 lexical choices impact lexical selection?

2- How does practice in rejecting L1-related distractors while learning L1-L2 incongruent collocations, relative to rejecting unrelated distractors, help regulate L1 interference?

3- Can experience with rejecting L1-related distractors aid in the suppression of interference from the native language during subsequent retrieval of L1-L2 incongruent collocations? Will learners who lack practice in suppressing L1-related distractors experience more L1 interference, and therefore produce more errors?

4- Will more strongly inhibited L1-related distractors be more costly to subsequently retrieve relative to unrelated distractors that do not require strong inhibition?
Predictions

Based on the literature, we put forward the following predictions:

1 – As shown in previous research on semantic interference (e.g., Schriefers, 1992), we predict that L1-related distractor verbs will induce cross-language interference, and result in selection costs, producing slower Reaction Times (RTs).

2 – Lexical selection in the L1-Interference group will require the engagement of enhanced monitoring and controlled selection (e.g., Rodriguez-Fornells, van der Lugt, Rotte, Britti, Heinze & Münte, 2005). We predict that the initial cost proposed in Prediction 1 (slower RTs in L1-Interference), will ultimately result in greater efficiency in selection (faster RTs in L1-Interference). This idea is in line with the seemingly paradoxical logic of desirable difficulties, which predicts that apparent initial costs and worse performance in more difficult learning conditions should transform into improved performance and results as learning unfolds.

3 - Practice of controlled retrieval should translate in better access to L2-specific representations in learners in the “L1-Interference” group, leading to enhanced learning relative to the “Unrelated” group. Learning will be measured in recall tests that will allow us to quantify the number of correctly recalled responses, as well as the number of errors that can be attributed to L1 interference (i.e., literal translations congruent with the L1 but not the L2).

4 – If, as predicted, practice in rejecting L1-equivalent distractor verbs helps suppress competition from the L1, those inhibited distractors should become harder to retrieve (slower RTs).
Method

Participants

A group of 49 learners of English was recruited at a university in Spain. Participants were native speakers of Spanish with low-intermediate proficiency in English (levels A2-B1 of the Common European Framework of Reference for Languages, 2011). All participants gave informed consent and were paid 10 USD per hour of participation. To confirm their eligibility, participants completed a baseline test; eight participants were excluded due to prior knowledge of the collocations. Additional details are provided in the “Testing” section. One participant who reported a learning disability was also excluded. The remaining participants (N = 40) were randomly assigned to one of two learning conditions: a group in which unrelated distractors were presented during the retrieval practice (henceforth, the “Unrelated” group; 75% female, 25% male), and a group that saw L1-related distractors during recall of incongruent collocations (“L1-Interference” group; 70% female, 30% male). The number of familiar items (Unrelated mean: 1.25; SD: 1.48; L1-Interference mean: 0.7; SD: 1.08) was not significantly different between groups ($t(34.8) = -1.3, p = .19$).

Individual differences measures

Participants were administered various measures to assess language proficiency in the L1 and L2, and to ascertain individual differences in cognitive control. Here we summarily enumerate all the measures, and report the results below (see Table 2-1 below). Complete descriptions of the tasks are available in Appendix A.

---

3 In determining the target sample size per condition, we aimed for a sample similar to that of other psycholinguistic studies that examined learning conditions under a controlled experimental environment in a laboratory setting (e.g., Cintrón-Valentin & Ellis, 2016; Finkbeiner & Nicole, 2003; Morgan-Short et al., 2014; Potts & Shanks, 2014), including experimental work focused on learning collocations (e.g., Pellicer-Sánchez, 2017; Sonbul & Schmitt, 2013); the sample size was, therefore, guided by the prior literature rather than by a power analysis.
To assess linguistic proficiency and background in the L1 and L2, participants completed an abridged version of the LEAP-Q (Marian, Blumenfeld & Kaushanskaya, 2007). General English proficiency was measured through an abridged version of the Michigan English Language Institute College English Test (MELICET). Participants also completed a multiple-choice test to assess knowledge of the individual words employed in the experiment, which they were expected to know.

Two measures of memory were collected: (a) A nonword repetition task adapted to Spanish phonotactics was used to measure Phonological Short Term Memory (PSTM) (Baddeley, Papagno & Vallar, 1988; Martin & Ellis, 2012); (b) Working memory was measured through a Spanish version of the Reading Span Task (Elosúa, Gutiérrez, García Madruga, Luque & Gárate, 1996).

Finally, to ensure that both groups were comparable in terms of their cognitive control abilities, participants completed the AX-Continuous Performance Task (AX-CPT) and the Flanker task (Eriksen & Eriksen, 1974).

Table 2-1: Summary of cognitive and proficiency measures for each group.

<table>
<thead>
<tr>
<th></th>
<th>Unrelated condition</th>
<th>L1-interference condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valid N</td>
<td>M</td>
</tr>
<tr>
<td>Age (in years)</td>
<td>22</td>
<td>24.05</td>
</tr>
<tr>
<td>Level of education (1-8)</td>
<td>22</td>
<td>4.4</td>
</tr>
<tr>
<td>OoA (in years)</td>
<td>22</td>
<td>4.6</td>
</tr>
<tr>
<td>MELICET (/50)</td>
<td>22</td>
<td>20.3</td>
</tr>
<tr>
<td>Weekly exposure to L2 (/1)</td>
<td>22</td>
<td>0.16</td>
</tr>
<tr>
<td>L2 Immersion (in years)</td>
<td>22</td>
<td>0.34</td>
</tr>
<tr>
<td>Eng. Picture Naming (accuracy %)</td>
<td>22</td>
<td>50</td>
</tr>
<tr>
<td>Span. Picture Naming (accuracy %)</td>
<td>22</td>
<td>95</td>
</tr>
<tr>
<td>PSTM: Nonword repetition (/1)</td>
<td>22</td>
<td>0.56</td>
</tr>
<tr>
<td>WM: Spanish reading span</td>
<td>22</td>
<td>50.35</td>
</tr>
<tr>
<td>AX-CPT (BSI)</td>
<td>22</td>
<td>0.40</td>
</tr>
<tr>
<td>Flanker effect (ms)</td>
<td>22</td>
<td>48.81</td>
</tr>
</tbody>
</table>

*Note*: Means and standard deviations are shown. Values represent raw scores, unless otherwise indicated.

Groups were matched across all but two measures. A significantly higher mean was found for the Unrelated group in the PSTM task ($t(37.6) = -2.3, p = .03$) as well as an earlier L2 Onset of Acquisition (OoA) ($t(32.1) = 2.43, p = .02$), relative to the L1-interference group.
Both these differences would predict improved learning ability for the Unrelated group relative to the L1-interference group, disfavoring our predictions.

**Materials**

Three types of materials were created: Materials for the Familiarization phase, for the Practice phase and for Testing. The materials for studying and testing were identical for all participants, while the materials used for practice differed across learning groups. All materials were based on the same list of 45 collocations.

For ease of presentation, we first describe the list of collocations used to create the experimental materials. Subsequently, in the section “Experimental Procedure”, we describe the materials and procedure for the Familiarization, Practice and Testing parts of the experiment.

**List of collocations**

Forty-five collocations were extracted using the web-based version of the Corpus of Contemporary American English (COCA, Davies, 2008) with over 560 million words, and data for equivalent Spanish collocations were extracted from the Corpus del Español (Davies, 2016)—with over 2 billion words. Three types of collocations were included (see Table 2-2 for a sample of the materials). The first type, which we will refer to as “L1-L2 incongruent”, differed in Spanish and English by virtue of the fact that the noun was equivalent across both languages but the verb was not. For example, the Spanish equivalent of “run a business” is “llevar un negocio”, which literally translates as ‘carry a business’. The second type contained “L1-L2 congruent” collocations derived from the first list. To create these, we took the verbs used in the first list (“run” in “run a business”), identified their idiomatic equivalents in Spanish (“llevar un negocio” – literally “carry a business”) and used the literal translation to create a new L1-L2 congruent collocation (“carry”; e.g., “carry [his] name”).
Each literally translated verb was combined with a new collocational noun that was selected from among the most frequent collocates found in COCA for that verb.

Finally, a third type of “semantically related collocations” also contained congruent collocations, but this time the verbs were semantically related to the verbs of the “incongruent” collocations (e.g., “walk the street”, where “walk” is related to “run”). This third type was included to test the ability of participants to learn collocations in which they would experience interference, but from a source different from the L1. At least one study has suggested learning costs when semantically related collocations are learned together (Webb & Kagimoto, 2011). In the present study the meaning of collocations is not related, but the difficulty stems from conflict between semantically related verbs (e.g., walk – run) during recall of the target verb for a given collocation that has been learned. In particular, this manipulation would allow us to examine whether more efficient interference suppression in learners in the L1-interference group might also allow for more efficient selection, despite potential within-language competition in the L2.

Words were not repeated across lists and were matched in word length, frequency, concreteness, and collocational strength. To determine the collocational strength of the verb-noun pairs, t-scores were used as the statistical association measure. The full stimuli are available in Appendix B.

<table>
<thead>
<tr>
<th>L1-L2 incongruent</th>
<th>L1-L2 congruent</th>
<th>Semantically related</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. English (L2)</td>
<td>run [a] business</td>
<td>carry [his] name</td>
</tr>
<tr>
<td>Spanish (L1)</td>
<td>llevar [un] negocio</td>
<td>llevar [su] nombre</td>
</tr>
<tr>
<td>Literal L1 translation</td>
<td>‘carry [a] business’</td>
<td>‘carry [his] name’</td>
</tr>
<tr>
<td>2. English (L2)</td>
<td>launder money</td>
<td>whiten [one’s] teeth</td>
</tr>
<tr>
<td>Spanish (L1)</td>
<td>blanquear dinero</td>
<td>blanquear [los] dientes</td>
</tr>
<tr>
<td>Literal L1 translation</td>
<td>‘whiten money’</td>
<td>‘whiten [one’s] teeth’</td>
</tr>
</tbody>
</table>
Experimental procedure

Data collection was done over a span of six weeks. In week 1, participants completed the first two sessions of the experiment. In Session 1, participants were administered all the proficiency and cognitive measures. This first session lasted about 2 hours. That same week, participants returned for Session 2. First, they completed the Familiarization phase (lasting approximately 20 min.); this was followed by the experimental Practice procedure (approx. 30 min.), which differed across conditions; and an immediate recall test.

The next two sessions, consisting of additional practice and tests, were completed the following week. The additional Practice sessions and tests allowed us to examine the effect of the manipulation after completing only one Practice session, and after three sessions.

Participants returned for short delayed tests in weeks 3 and 6. Delayed tests were considered critical to assess retention beyond immediate tests, in line with real-life goals of language learning. They should provide a more reliable measure of the long-term impact of the experimental manipulation and the durability of any significant effects observed in immediate testing. An outline of the data collection protocol is presented in Table 2-3.

Familiarization phase

Familiarization materials

Familiarization materials were created for auditory and visual presentation of the forty-five English collocations and their Spanish equivalents. The English collocations were recorded by a native English speaker. Spanish collocations were recorded by the first author, a native speaker of the same variety as the speakers tested.
**Familiarization procedure**

In the Familiarization phase, participants studied the list of 45 collocations, which were presented simultaneously auditorily and visually on a computer screen using E-Prime 2.0 (Schneider, Eschman & Zuccolotto, 2002). This phase served the purpose of familiarizing learners with the target collocations, which would then be practiced under different experimental conditions for each group in the subsequent Practice procedure. Visual presentation of the collocations was done as follows. First, the Spanish collocation was displayed in the center of the screen, followed by a second screen with the English collocation. Participants were then prompted to repeat the English collocation in two subsequent screens, first by saying it out loud, and then by typing it. After a collocation was typed, feedback was provided by showing the collocation in blue or red. If participants typed the collocation exactly as it was shown, the collocation was displayed in blue font; otherwise, the correct collocation appeared in red font (see Figure 2-1). Overt vocal repetition was used because it has been suggested that it aids learning (e.g., Ghazi-Saidi & Ansaldo, 2017); typed responses allowed to provide automatized feedback during this phase.

Because seeing each of the 45 collocations only once might not provide sufficient familiarization, a repetition round was administered. This time, in addition to repeating the same procedure, participants were asked to recall the Spanish equivalent of the collocations. This was considered important so that learners would not simply focus on learning the form of L2 collocations, while ignoring or failing to remember their meanings. Given the length of the full list (with a total of 45 items), the list was broken down into five segments, with meanings being recalled in blocks of nine collocations.
Oral responses were recorded. Groups of participants did not differ in the ability to recall the meanings of collocations (Unrelated: 85.8%, SD: 35.0; L1-Interference: 85.3%, SD: 35.4; $t(1797.7) = -0.3$, $p = .79$). This Familiarization phase was followed by the Practice procedure. The same Practice procedure was also completed in Sessions 3 and 4 in Week 2, as shown in Table 2-3. The additional Practice sessions conducted in Week 2 allowed us to examine the effect of the manipulation after completing only one Practice session, and after three sessions at the end of the second week.

Table 2-3: Testing protocol.

<table>
<thead>
<tr>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>Session 2</td>
<td>Session 3</td>
<td>Session 4</td>
</tr>
<tr>
<td>Tasks</td>
<td>Cognitive battery and proficiency measures</td>
<td>Familiarization and Practice</td>
<td>Practice</td>
</tr>
<tr>
<td></td>
<td>Immediate, after practice</td>
<td>Delayed, at beginning</td>
<td>Immediate, at end of session</td>
</tr>
<tr>
<td>Tests</td>
<td>No practice</td>
<td>1-month delayed test</td>
<td>1-month delayed test</td>
</tr>
</tbody>
</table>
Practice

Practice materials

The materials for the Practice procedure were based on the forty-five collocations studied (e.g., “run a business”). The Practice consisted of forced-choice trials in which the target verb of a collocation had to be selected. Each trial contained two verbs (the target, e.g., “run”; and a distractor verb, e.g., “touch”), and the associated noun in a collocation. In the two experimental learning groups, trials sequentially presented two verbs followed by a noun (V1-V2-N, e.g., “run – touch – business”).

One list of Practice materials was created for each of the two experimental learning groups. The two lists differed only in the distractor verbs presented for trials with incongruent collocations. In each list, the incongruent collocations (e.g., “run a business”) were associated with distractor verbs specific to each learning group. One group saw unrelated distractors (henceforth the ‘Unrelated’ group) and another one saw distractors that would be congruent with the native language (‘L1-Interference’ group). For the previous example (“run a business”), the ‘Unrelated’ distractor was “touch”, while the ‘L1-Interference’ distractor was “carry”; to reiterate, “carry” was the literal translation of the verb in the Spanish equivalent, as shown in Table 2-2 above. To illustrate, participants in the ‘Unrelated’ group saw e.g., run – touch – business (correct response “run”). For the same collocation, participants in the ‘L1-Interference’ group had to correctly discard the L1-equivalent distractor verb in e.g., run – carry – business.

The three words in each trial (e.g., run – carry – business) were presented sequentially. The order of target and distractor was counterbalanced, so that the distractor would appear first half of the time, and the target first in the other half. Importantly, all verbs (targets and distractors) are potential candidates, as all are part of the familiarized collocations (i.e., “carry” is part of “carry his name” and “run” appears in “run a business”). Therefore, having seen “run” and “carry” it is not until “business” is displayed that learners could know
what verb to select. Each collocation appeared 9 times per list, producing a total of 405 trials per participant in each session.

**Practice procedure**

Participants were presented with two verbs followed by a noun (V1–V2–N), and were required to respond orally by selecting the appropriate verb accompanying the noun; that is, participants responded by pronouncing the verb selected only. The RTs of the onset of oral responses were automatically recorded by a microphone connected to an SR Box (Psychology Software Tools, Inc.), and accuracy was coded offline. In each trial, a fixation cross was first displayed for 1 second. Then each verb was presented sequentially for 700 ms. The noun was presented for 300 ms, followed by a fixation cross shown for up to 6 seconds or until an oral response was registered, whichever came first. A longer presentation time of the verbs, rather than the noun, was used so that participants would attend to the verbs, and only rely on the noun as a cue for selection between candidates. Feedback was provided by presenting the correct verb for 500 ms after each trial.

**Testing**

**Testing materials**

First, a baseline multiple-choice (MC) test was created in order to assess any potential familiarity with the collocations to be learned, and with the goal of excluding participants with prior knowledge of the target materials. Example 1 below presents a sample item. Each item presented four choices. For the critical incongruent collocations, the four choices contained the target verb (e.g., *run*), the non-target literal Spanish equivalent (*carry*), the associated semantically related verb (*walk*) and a fourth verb (*bring*) in randomized order.
(1) llevar un negocio - _____________ a business
(a) run (b) carry (c) walk (d) bring

Participants were asked to provide confidence ratings using a scale from 1 to 5 (1 = no knowledge; 5 = certainty in the response), and those who indicated previous knowledge of at least one collocation (rating = 5) or substantial familiarity with more than three (rating ≥ 3) were excluded (N = 8).

A L1-to-L2 translation test was created to assess immediate and delayed recall at different points throughout the study. For each question, the same Spanish meanings of the English collocations used for the Familiarization phase were presented. Tests were administered in order to assess learning (immediate tests) and retention (delayed tests).

The MC recognition test was used as a baseline rather than a production study, because the inability to recognize the correct verb is a more stringent test of null familiarity than the inability to produce a valid response. That is, our expectation was that learners would not be able to recognize the correct choice when presented to them.

Testing procedure

Learners completed immediate tests at the end of the first (Week 1) and last (Week 2) sessions of Practice. The first of three delayed tests was completed in Week 2, while the experiment was still ongoing. Participants returned approximately one week after Practice 1 (mean: 6.2 days; SD: 0.94), and completed the first delayed test before doing additional Practice. The two remaining delayed tests were administered once all Practice was completed. The second and third delayed tests took place approximately one week (mean: 6.8; SD: 1.98) and one month (mean: 31.97; SD: 6.27) after the last Practice session, respectively. One participant in the Interference group did not return for the one-month delayed test; her results are included for all other parts of the study.
For each of the 45 collocations, a screen was presented with a Spanish verb-noun sequence and participants were asked to recall and type the equivalent L2 English collocation. For the immediate tests and the first delayed test, feedback was provided after each response by presenting the verb-noun for 1,500 ms. Since the training procedure involved only recall of verb-noun associations but not of their meaning, feedback on the tests allowed participants to check their form-meaning representations once they had responded. No feedback was provided in the one-week and one-month delayed tests. Responses were coded offline for accuracy.

**Results of recall tests**

In this section, we first present the results of the recall accuracy for the immediate and delayed tests that learners completed. This analysis allowed us to address RQ 3, that is, we examine whether experience rejecting L1-related distractors (L1-Interference group) will result in fewer errors when retrieving L1-L2 incongruent collocations. To further address the question of whether learners in the ‘Unrelated’ group would experience more interference from the L1, we also conducted an analysis of the types of errors. In the “Reaction Time analysis” section, we report RT data for responses in each of the three training sessions.

Typed responses to recall tests were coded for accuracy based on a 2-letter rule. That is, partially correct spellings were accepted as long as no more than 2 characters were misspelled and the correct target word could be identified.

**Accuracy in recall**

The results of the recall tests were analyzed using mixed-effects logistic regression. This type of analysis is ideal for binary dependent variables (i.e., accurate or inaccurate response), as they allow to analyze the unaggregated data rather than means (Jaeger, 2008).
All analyses reported in this and other sections were carried out with the lme4 package (Bates, Maechler, Bolder & Walker, 2015) in R version 3.3.2 (R Core Team, 2016).

The analysis examined learning across groups as measured by the tests administered throughout the study (see Figure 2-2). Following attempts to build maximally specified models (Barr, Levy, Scheepers & Tily, 2013), which led to convergence issues, the random effects structure was simplified (Bates, Kliegl, Vasishth & Baayen, 2015). Final models included random intercepts for subjects and items (Baayen, Davidson & Bates, 2008). In order to control for the baseline differences between groups reported above, models also included by-item slopes for individual PSTM. Fixed effect factors included Group (Unrelated or L1-Interference distractors), Type of collocation (congruent, incongruent or semantically related), Test (each test, 5 levels) and their interaction. The contribution of cognitive control was investigated by including fixed effects for the Behavioral Shift Index (from the AX-CPT), and Flanker Effect. Individual measures of memory were also considered: Nonword repetition (PSTM), and Reading Span (WM). All continuous variables were centered (Baayen et al., 2008).

For each sublist of collocations, models were built starting with a simple mixed-effects structure with Group, Type of collocation, Test, and their interaction, as predictors. Due to convergence issues in models containing the three-way interaction of Group x Type x Test, two-way interactions for Group x Type, Group x Test, and Type x Test were included. In a step-by-step forward model selection procedure, predictors and their interaction with Group and Type were introduced one by one, and were kept if the model fit was significantly improved (likelihood ratio test, $p < .05$). The reference levels were set to Unrelated for Group, and Congruent for Type of collocations. The results are reported below; the model output is available in Appendix C. Parameter-specific $p$–values were estimated using the normal approximation. Figure 2-2 shows the progression of performance across groups at each test and for each collocation type.
The analysis revealed a main effect of collocation Type, with significantly lower recall for incongruent collocations in all tests ($\beta$: -2.03, SE: 0.58, $p < .001$). There was a marginally significant main effect of Group, with higher accuracy in the 'L1-Interference' group ($\beta$: 0.8, SE: 0.45, $p = .07$). Critically, the interaction of L1-Interference Group and incongruent Type was highly significant ($\beta$: 0.81, SE: 0.2, $p < .0001$), showing that learners in the 'L1-Interference' group had higher recall rates for incongruent collocations.

The results confirm the pattern shown across testing in Figure 2-2. Relative to the first immediate test, overall accuracy was significantly lower in the first delayed test ($\beta$: -0.62, SE: 0.23, $p < .01$), but higher for tests completed after additional practice (all at least $p < .05$).

However, the interaction of Test x Type showed lower accuracy for incongruent collocations in the Delayed test 1, after only one practice session ($\beta$: -0.92, SE: 0.26, $p < .001$); and in the one-month delayed test ($\beta$: -0.9, SE:2.28, $p < .01$). Similarly, the interaction of Test x Group pointed at a reduced effect of Group for Delayed tests 1 (one week after Practice 1; $\beta$: -0.58, SE: 0.21, $p < .01$) and 3 (one month after Practice 3; $\beta$: -0.49, SE: 0.23, $p < .05$). To further investigate these interactions, we conducted dedicated follow-up analyses with the same methodology described above. The results showed no effect of condition in Delayed test 1, but a marginally significant interaction of Type in the one-month delayed test.
after completing additional practice ($\beta$: 0.8, SE: 0.44, $p = .07$). These results, altogether with the significant interaction in Delayed 2, revealed that differences in retention were persistent only after practicing recall in the three Practice sessions.

**Types of errors**

To fully address RQ 3 and explore potential evidence of L1 interference during recall, we conducted an analysis of the types of errors conducted by learners in each group. Specifically, we aimed to explore if learners in the ‘Unrelated’ group used literal L1 translations more than learners in the ‘L1-Interference’ group (who practiced suppression of L1-related translations). In the analysis of types of errors, incorrect test responses that contained an equivalent of the L1 verb (e.g., “carry a business” was produced instead of “run a business”) were considered “calques”; the proportion of these was compared against any other type of non-target response (including lack of response). In the Unrelated group, calques accounted for 36% of the errors, but in the L1-Interference group they were only 22% of all errors. A t-test revealed this was a statistically significant difference ($t(292.92) = 2.73, p < .01$). The error-type analysis further supports the hypothesis that inducing interference during practice afforded protection against interference in retrieval.

**Reaction time analysis**

This section presents the results of the RT data of responses to trials in the training procedure described above. We first present the results of a Growth Curve Analysis (GCA) on the RTs of incongruent collocations for each individual session, to address RQ 1 and 2 and predictions 1 and 2 formulated above. Questions 1 and 2 were theoretically motivated by the research on desirable difficulties in learning, which predicts that initial costs will transform into advantages. RQ 1 asks whether L1-related distractors will initially cause a cost in selection, i.e., slower RTs; RQ 2 asks whether experience in resolving conflict in responses
leads to a more efficient selection process as learning unfolds, relative to the condition with low-conflict (‘Unrelated’) distractors. We then report the results of a by-group comparison of RTs for different types of collocations in each training session, to test RQ 4 and prediction 4.

All analyses were performed on z-score normalized RTs. Trials with invalid RTs due to microphone failure (2%) were removed. Incorrect responses were removed based on offline coding (Unrelated: 7%; L1-Interference: 6%). Responses with RTs shorter than 400 ms after presentation of the noun, or longer than 5,000 ms were excluded (Unrelated: 3%; Interference: 3%).

**Growth curve analysis of reaction time**

Growth curve analysis (GCA; Mirman, 2014) was used to analyze the RTs in selection of verbs for incongruent collocations over the course of each of the three Practice sessions. By-session analyses were necessary to determine the shape of the curve within each individual session.

The overall learning curves were modeled with second-order orthogonal polynomials and fixed effects of Group (‘Unrelated’ or ‘L1-Interference’) on all time terms. Since the goal of this analysis was to examine selection of the verb only in the incongruent collocations, Type of collocation was not included. Trial order within the training session served as the time dimension. The unrelated group was treated as the baseline and parameters were estimated for the interference group. The model also included random effects of participants on all time terms, as well as by-item random slopes for Group, PSTM and OoA. The same fixed effects and interactions as in the accuracy analysis were considered (i.e., PSTM, Flanker Effect, BSI, WM), were added individually and their effects on model fits were evaluated.

---

4 Analyses were done by-session, rather than for all at once with session (i.e., time) as a variable. This was done because the significance of some effects might vary from one session to another; by-session analyses helped us avoid potential three- and four-way interactions and allow for easier interpretation of the results.
using model comparisons. Parameter-specific p-values were estimated using the normal approximation.

For ease of visual interpretation, the data were binned, as is common practice in GCA (Mirman, 2014, p. 20). In the present analysis, data were grouped into 15 bins; this number of bins maximized the proportion of trials per bin within a range of 10-20 bins. Figure 2-3 shows the RT data and model fits for the three sessions of training. The model output is available in Appendix C.

![Figure 2-3: Growth Curve Analysis of RTs for verb selection across the three practice sessions. For the second training, one single solid line represents no significant differences across conditions. All RTs used in the analysis and figures are z-scored. Error bars represent the standard error.](image)

**Practice session 1**

The effect of Group on the intercept did not improve model fit ($\chi^2 (1) = 0.22; p = .64$), nor did the effect of Group on the quadratic term ($\chi^2 (1) = 0.95; p = .33$). The effect of Group on the linear term, however, improved the model fit ($\chi^2 (1) = 6.06; p < .05$). The interaction of Group with the linear term indicated that the two training conditions differed in the rate of learning, with faster learning in the L1-Interference group ($\beta: -1.09$, SE: 0.42, $p < .01$). The results revealed an effect of PSTM ($\beta: -0.05$, SE: 0.02, $p < .01$).
**Practice session 2**

In the second practice session, Group had no effect on the intercept ($\chi^2 (1) = 1.45; p = .23$), nor on the linear ($\chi^2 (1) = 0.26; p = .61$) or quadratic terms ($\chi^2 (1) = 1.94; p = .16$), suggesting no significant differences across groups.

**Practice session 3**

No effect of Group was found on the linear ($\chi^2 (1) = 1.27; p = .26$) or quadratic terms ($\chi^2 (1) = 1.32; p = .25$). The model was improved by the effect of Group on the intercept ($\chi^2 (1) = 4.11; p < .05$), showing overall shorter RTs in the interference group in the session ($\beta: -0.1, SE: 0.04, p < .05$).

**Reaction times for verb selection across sublists of collocations**

In a final analysis of the RT data, we investigated the cost of selecting the verb for the different lists of collocations across the two learning groups. This allowed us to address RQ 4 and test the associated prediction: that the need for greater inhibition should be associated with a greater cost in retrieval. A mixed-effects linear regression analysis was performed using the same software and procedures described above. In contrast with the GCA reported above, the linear regression analyzed differences within each whole training session. This means that it cannot reveal changes that occur over the course of one session; rather, here we analyze differences in the averages between Groups (Unrelated, L1-Interference) and Types of collocations (Incongruent, Congruent, Semantically related) for the entire session.

Subjects and items were included as random intercepts. The same predictors as in the generalized mixed-effect regression were considered. Final models included by-item random slopes for OoA. Figure 2-4 shows the RTs of responses in each group for all Practice sessions. Further details are provided in Appendix C.
In comparison with the GCA, models predicting RTs for Sessions 1 and 2 were not improved by adding Group (Session 1: $\chi^2 (1) = 0.03; p = .86$; Session 2: $\chi^2 (2) = 0.83; p = .66$), nor Type of collocation (Session 1: $\chi^2 (2) = 0.18; p = .91$; Session 2: $\chi^2 (1) = 0.01; p = .92$) nor their interaction (Session 1: $\chi^2 (5) = 1.72; p = .89$; Session 2: $\chi^2 (5) = 3.67; p = .6$).

However, a significant interaction of Group with Type of collocation was revealed for Session 3, showing that while RTs for verb selection did not differ for congruent collocations across groups, responses for incongruent collocations were faster in the L1-Interference group ($\beta: -.01$, SE: 0.04, $p < .05$). This result supported our prediction; we discuss the implications of this result in the following section.

There was an interaction between the Flanker effect and collocation type only in Session 1, but not in subsequent sessions. A larger Flanker effect was associated with faster RTs in the incongruent ($\beta: -.05$, SE: 0.02, $p < .01$) and semantically related ($\beta: -.06$, SE: 0.02, $p < .01$) lists. Finally, a significant interaction of PSTM with type of collocation

---

**Figure 2-4**: Whole session RT averages for each of the collocation sublists. The top figure shows RTs in the group of learners with unrelated distractors, and the bottom figure shows the group in the L1-interference condition. All RTs used in the analysis and figures are z-scored. Error bars represent 95% CIs.
emerged for all three sessions, revealing that greater PSTM facilitated selection only in the incongruent and semantically related lists.

We conducted a post-hoc analysis to investigate whether, at the individual level, higher learning gains were associated with L1 inhibition. As an index of interference, we calculated the difference of verb selection RTs in congruent trials minus incongruent trials for the last Practice session (with more positive values indicating inhibition), and correlated it with the scores of the one-month delayed test (Figure 2-5). A significant correlation for the Unrelated group \((r_s = .59, p < .01)\), confirmed that learners that showed greater inhibition in the L1 were those who tended to have higher recall rates one month later. The analysis of the Interference group, in which L1 inhibition and retention were highest and more generalized, revealed no significant correlation \((r_s = .15, p = .53)\).

![Figure 2-5: Recall accuracy rates are for the one-month delayed test. The index of L1 inhibition represents the cost of selecting L1-congruent verbs (RTs of L1-equivalent verbs minus RTs of L1-incongruent verbs) in the last practice session. Positive values indicate inhibition of L1-equivalent verbs.](image)

**Discussion**

Languages differ in how they encode frequently expressed concepts into conventionalized collocations. Therefore, acquiring a new language also involves learning to
combine individual words in ways that are recognizable to other speakers, not only in terms of their syntactic structure, but also in their meaning. For adult L2 learners, entrenched knowledge of words and conventional collocations in their native language (e.g., *run a business*) interferes when L2 sequences that are incongruent with those representations must be learned (e.g., *llevar un negocio*, literally ‘carry a business’ but equivalent to “run a business”; Nesselhauf, 2003; Boers, et al., 2014; Peters, 2016). Learners differ in their ability to establish, and then retrieve, these L1-L2 incongruent collocations. We compared learning in two groups to examine if the ability to inhibit competing L1 representations during learning can be trained, and whether this kind of regulation may enhance learning and recall of L2 incongruent collocations. The degree of L1 interference was manipulated in a forced-choice retrieval practice procedure by presenting distractors that were either related or unrelated to the native language equivalents of the sequences being learned.

In a recent study on individual word learning, Bogulski and colleagues (2018) proposed the L1 Regulation Hypothesis, suggesting that learning through L1 translations would engage mechanisms to regulate the native language. The study, which included different groups of bilinguals, demonstrated that those learning through their L1 adopted the strategy of taking longer study time and showed a learning advantage. However, the study did not provide evidence of inhibition of the L1, critical to the regulation hypothesis. The results of the current study showed for the first time that learning conditions that involve the L1, even if implicitly, result in native language inhibition associated with enhanced learning.

---

5 These findings suggest certain advantages in using the native language to develop L1-to-L2 mappings and L1 regulation, and therefore encourages further research into how the L1 may bootstrap L2 learning. The specific pedagogical implications are beyond the scope of this paper. However, recent influential lines of research claim a role for the native language in second language learning in different contexts (e.g., Cummins, 2000) in ways that are distant from the grammar translation methods. The potential benefits of using the L1 in particular contexts are in line with contemporary multilingual pedagogical approaches such as Translanguaging (e.g., García & Wei, 2014; García, Johnson, Seltzer, & Valdés, 2017).
Enhanced learning of incongruent collocations

In line with previous studies (e.g., Peters, 2016), the results of all immediate and delayed recall tests showed that accuracy (i.e., the proportion of correctly recalled items) was significantly lower for incongruent than for congruent collocations in both groups. While this was true for both learning groups, the comparison of recall rates in each group (L1-Interference and Unrelated) revealed that our experimental manipulation was successful in enhancing learning of incongruent collocations in the experimental group (L1-Interference) relative to the baseline (Unrelated). As predicted, inducing interference through L1-related distractors produced significantly greater accuracy for incongruent collocations, even after one single practice session, suggesting that this type of conflict was a desirable difficulty in learning. Further, the testing scheme employed allowed to assess the amount of practice needed. The results of the first delayed test showed that, after completion of one practice session, between-group differences did not persist when participants were tested one week later. In the second delayed test, one week after two additional practice sessions, recall rates (i.e., the proportion of correctly produced collocations in the L2) remained unchanged. In other words, completing only one practice session produced transient learning, but three practice sessions resulted in stable recall rates. Additionally, the one-month delayed test examined the stability of learning for L2 collocations beyond the two-week period commonly measured in other studies on learning of collocations (e.g., Sonbul & Schmidt, 2013; Boers et al., 2014). When learners were tested again after one month without practice, a marginally significant advantage was still found in accuracy for the L1-interference group. The test results thus showed that practice in the “desirably difficult” condition, in which L1-related interference was induced, led to increased accuracy in recall of the critical incongruent collocations.
Native language regulation as part of L2 learning

While the results of the recall tests support our predictions, a main goal of this study was to examine the mechanisms conducive to enhanced recall of collocations. We proposed that induced L1-interference during recall is a desirable difficulty in learning (Bjork & Kroll, 2015; Bogulski, Bice & Kroll, 2018), and therefore that initial costs should later lead to greater efficiency in suppressing irrelevant distractors during recall. Specifically, we predicted that if learners who were trained in the L1-Interference condition became more efficient at target selection in incongruent collocations, this should be observable through increasingly faster RTs, and a more rapid rate of learning incongruent collocations. The GCA results were fully consistent with this prediction. The interaction of Group and Type of collocation for RTs in Practice session 1 showed that learners in the L1-Interference group became faster at a faster rate in selecting the verbs of L1-L2 incongruent collocations than learners in the Unrelated group. The comparison of both groups in Practice session 1 (in Figure 2-3) shows that initially slower RTs in the L1-Interference group –indicating a greater cost in selection– translated into faster RTs by the end of the training session, relative to the Unrelated group. By Practice session 3, the GCA revealed that RTs in the L1-Interference group were faster throughout the entire session. That is, harder retrieval conditions in the Interference group led to faster performance than in the Unrelated group during verb selection for incongruent collocations. This supports our claim that conflict-inducing learning conditions engaged control mechanisms that contributed to more efficient selection of incongruent collocations.

An alternative explanation worth considering is that between-group differences were not due to the engagement of mechanisms such as conflict-monitoring and inhibition. Rather, it is possible that presenting L1-related distractors during practice allowed some learners to establish more direct connections between L1 and L2 equivalents. Seeing the verbs of incongruent collocations along with the functionally equivalent verbs of the L1 might have strengthened their association. This interpretation might also be supported by the Depth-of-
Processing Hypothesis (e.g., Craik & Lockhart, 1972; Craik & Tulving, 1975), in the sense that seeing L1 and L2 equivalents might have encouraged a deeper level of processing and perhaps comparison (whether consciously or not) between the form and meaning of the collocations across the two languages. Indeed, establishing a more direct association might have helped learners establish a form-meaning mapping across languages. Some models of L2 access, such as the Revised Hierarchical Model (Kroll & Stewart, 1994), posit an imbalance in the directionality of lexical connections, with L2-to-L1 connections being weaker than L1-to-L2. Therefore, creating stronger links between the entrenched L1 lexicon and weaker L2 lexicon might aid learning.

While this interpretation cannot be excluded, there are two important points that would downplay the role of such a direct connection in our results. First, accuracy in selection during practice was extremely high in both the Unrelated and the L1-Interference group, indicating that learners in neither group experienced any difficulty recognizing and selecting the target verbs. Therefore, rather than viewing the representation of the English collocations as problematic, the results strongly point to an interpretation in which cost experienced in retrieving these novel L2 representations is the critical factor. In other words, in the present paradigm, there is no direct evidence that the locus of difficulty is establishing a representation, nor that different input was required by the Unrelated group in this respect; rather, it is retrieval that seems problematic. A second crucial consideration to note is that the implications of strongly associated L2 and L1-equivalent verbs would run counter to the RT data in our results: greater co-activation of equivalent L1 and L2 forms would not result in less competition but, quite on the contrary, in greater competition from the L1 counterparts. It might well be the case that an association is formed between the target and non-target (i.e., distractor) candidates. Stronger co-activation would result in slower RTs due to conflict in selection; unless the selection system exerts inhibition on the distractors, alleviating competition from highly active competitors. Therefore, while exposure to verb pairs might facilitate associative links (note that this would apply not just to incongruent collocations, but to all trials), it is unclear how such potential associations might benefit, rather than hamper,
conflict resolution among competitors. The only explanation compatible with the finding of faster rejection of distractors in the Interference group, is that selection among competing representations occurs in parallel with increasingly efficient suppression of the non-target candidate.

**Evidence of L1 inhibition during learning**

Crucially, the specific prediction that the competitor verbs of incongruent collocations would become inhibited during learning was supported. While both groups of participants performed at ceiling in recalling the verbs of congruent collocations (as expected, given that these congruent collocations are presumed to rely on the L1), emergent costs in retrieval where apparent in the RTs. The results revealed that selection of the verbs of congruent collocations, which were easiest to recall before training (e.g., carry in carry her name), became slower after learning the incongruent collocations that require not selecting those same verbs (e.g., learning run a business requires not selecting the L1-compatible choice carry a business). In other words, the verbs that were non-targets for incongruent collocations (e.g., carry) still had to be selected on some trials as valid targets for congruent collocations; but inhibiting those verbs in incongruent trials (i.e., when they were presented as distractors) resulted in slower RTs for congruent trials where they did need to be selected (i.e., they were the target). This pattern provides strong evidence for the expected inhibition in a specific subset of verbs.

Another piece of evidence supporting the association between L1 regulation and learning comes from within-group variability in the Unrelated condition. While learners in that group had, on average, lower recall rates, some participants did attain high learning outcomes. A significant correlation for the Unrelated group, confirmed that learners that showed greater inhibition in the L1 in Practice session 3 were those who tended to have higher recall rates one month later.
These findings point at an association between recall conditions and long-term retention, and suggest that linguistic units that are susceptible to interference during retrieval (potentially resulting in difficult or failed retrieval) benefit from practice conditions that aid development of the necessary interference suppression skills. This idea is the core of theories of learning that posit advantages stemming from conditions that pose desirable difficulties. While the results of recall after one month suggest potential benefits in long-term retention, it is likely that, without any further exposure, gradual memory decay will have a greater long-term impact that outweighs the advantages conferred by more efficient language regulation. Further research will be needed to explore the interaction between language regulation and gradual memory decay in long-term retention (e.g., following months or even years).

Regulation of lexical competition within the L2

An additional, secondary finding was that not only were the L1-equivalent verbs inhibited, but also the verbs of the congruent collocations in the semantically related sublist (e.g., walk is related to run), as shown in Figure 2-4. This should not be surprising, given that within-language inhibition is even more straightforwardly accounted for, fully consistent with the notion that hard-to-retrieve words require inhibition of more highly active competitors. That is, selection of the L1-incongruent verbs resulted not only in the expected inhibition of L1-equivalent verbs (carry – run), but also in inhibition of within-language semantically related competitors (walk – run). Given the difficulty in selection of verbs in L1-incongruent collocations, more highly active within-language semantically related competitors must also be suppressed. This pattern of results provides insight into effects of both cross- and within-language inhibition in bilinguals.

\[^6\] One might wonder about the potential consequences of L1 inhibition for attrition of the native language. While this point will need to be addressed in further research, we would like to argue that language attrition is not necessarily only by L1 inhibition, and requires also decreased usage of the L1.
Finally, a comment regarding competition between different representations is in order. Our stimuli contained collocations along a continuum from more compositional to more idiomatic. As an anonymous reviewer pointed out, the idiom processing literature suggests that idiomatic strings, once acquired, result in significantly faster processing than non-idiomatic strings (e.g., Cacciari & Tabossi, 1988; Libben & Titone, 2008; Titone & Connine, 1999; Titone & Libben, 2014). Incongruent collocations tend to be by definition less cross-linguistically congruent and more idiomatic. Therefore, while incongruent collocations tend to be processed more slowly, less compositional incongruent collocations might show less of a disadvantage. On the other hand, non-native speakers do not always show the same advantages in processing idioms as native speakers (Siyanova-Chanturia et al., 2011), although this may change as proficiency develops (Yeganehjoo & Thai, 2012). However, our materials are not able to directly test this question.

Types of errors

To further explore potential evidence of inhibition from a different angle, we examined the types of errors produced by learners in the recall tests. A straightforward assumption is that interference from the native language will result in more errors due to use of L1-equivalent verbs (known as “calques”). In this sense, two opposite predictions can be made. The main hypothesis proposed here is that presenting L1-like verb distractors allowed learners to regulate competition from the L1. If this explanation is behind the higher learning rates in the L1-Interference group, we should see, on the one hand, low rates of errors due to L1 intrusions in that condition. On the other hand, learners in the Unrelated group—even though they were not presented with L1-related choices—would be predicted to produce more calques. Such a pattern would be further evidence that errors are due to the lack of experience in suppressing interference.

The alternative approach, discussed above, is that presenting the L1-equivalent verbs of incongruent collocations simply reinforced L1-L2 associations, and better recall was not
due to better monitoring of interfering representations, but to facilitating cross-language associations. This would lead to very different predictions: because learners in the Unrelated group never saw the L1-congruent verb paired with the L2-target verb, L1-based calques should be rare. On the other hand, learners in the L1-Interference group, who repeatedly saw the L1 equivalent, might produce a high number of L1-congruent intrusions. The error-type analysis reported provides further support to the hypothesis that inducing interference during practice afforded protection against interference in retrieval.

**Individual differences in memory**

The role of two different memory constructs—namely, PSTM and WM—was examined by including these measures as fixed effects in the analyses performed. The results showed the paramount role of phonological memory in learning of collocations, in the line of similar findings in previous studies of single word learning (Kaushanskaya, 2012; Martin & Ellis, 2012). The main effect of PSTM in every recall test points to the unique contribution of phonological memory to the encoding and retrieval of collocations as unitary chunks. WM was only significant in the second immediate test, administered after all three sessions of retrieval practice were completed, but not in the other tests.

It is worth noting that the second immediate test was completed at the point in which the benefits of practice were maximal. The significant contribution of WM in this particular test suggests an important role in allowing retrieval of the correct verb-noun combinations at a time in which all competitors were maximally active. We suggest that the ability to resolve competition among simultaneously active candidates may be the mechanism behind the significant contribution of WM.
The role of cognitive control in learning incongruent collocations

The contribution of cognitive control was investigated by considering the measures from Flanker task and AX-CPT in the analyses. Because retrieval of target L2 collocations was dependent on inhibition of L1 competitors, we predicted a significant role of cognitive control. The results showed that the Flanker effect scores (RTs in incongruent trials – RT in congruent trials) were predictive of RT group averages in the first practice session. Somewhat counterintuitively, a greater Flanker effect (indicating less efficient inhibition) was associated with faster RTs in selection of the target verbs in incongruent collocations during retrieval practice 1 ($β$: -0.05, SE: 0.02, $p < .01$); no effects were found for practice 2 and 3.

As discussed, better learning was associated with a greater initial cost in selection, shown by longer RTs in Practice 1 (see Figure 2-4). That is, greater learning was associated with slower RTs, which reflected that conflict-monitoring mechanisms were in fact being engaged appropriately. In less successful learners, RTs were faster, as shown by the results for the Unrelated group. Thus, the finding that those learners with greater ability to detect conflict experienced greater delays in response selection is consistent with the notion that enhanced learning was associated with a greater initial cost, and that experiencing greater interference was associated with greater gains when L2 representations are in conflict with the L1.

Generalizability of efficient learning

In addition to investigating the effect of distractors on recall of the incongruent collocations, we asked whether greater ability to select among competitors in some learners might translate in an advantage when resolving other types of interference. To this end, a subgroup of the congruent collocations was included, whose verbs were semantically related to the verbs of the incongruent collocations (e.g., run – walk). A greater cost of selection in this sublist of congruent collocations was observed in a number of ways. First, the analysis of
whole session RTs showed that individual PSTM predicted speed in verb selection not just for the incongruent collocations but also for collocations with semantically related verbs. Moreover, in Practice 1, individual Flanker effect scores also predicted RTs in verb selection for the incongruent and semantically related sublists, but not for the other congruent collocations. Despite the cost from semantic interference reflected in RTs, the lack of interactions suggested there were no differences across groups. A limitation to be noted is that recall of congruent collocations (including the semantically related) produced ceiling effects. Future studies should further investigate the potential spillover from enhanced efficiency in learning by examining more challenging learning conditions that prevent ceiling effects.

Finally, we can speculate that a desirable difficulties-based approach might confer other advantages in learning. However, desirable difficulties in learning e.g., tense morphology, may focus on mechanisms other than interference. For instance, learning of morphological cues specific to the L2 may require avoiding reliance on known cues that also exist in the L1 (e.g., see Cintrón-Valentín & Ellis, 2016; Ellis et al., 2012).

**Conclusion**

The current study provides new evidence of the weight that regulating competition from the native language carries in learning collocations in another language. Moreover, it tested a new practice paradigm focused on suppression of L1-interference as a new type of desirable difficulty in language learning. The results provided evidence that retrieval practice in interference-induced conditions enhanced learning gains of incongruent collocations relative to the baseline (unrelated-distractor) group. Further, the association between L1-inhibition and L2 learning gives critical support to the L1 Regulation hypothesis (Bogulski, Bice & Kroll, 2018). Future research should further investigate how engaging cognitive processes through particular learning conditions may produce not just better learning outcomes but also better learners. The findings reported also have practical implications for learning and teaching pedagogy, and for the design of learner-oriented tests and materials.
Chapter 3

The neural correlates of conflict detection and resolution during multiword lexical selection: Evidence from bilinguals and monolinguals

Introduction

This second experiment aimed to investigate the mechanisms of conflict detection among competing lexical choices. This study used a novel Stroop-like task to examine lexical selection in multiword units, where a word to be selected is strongly associated with an ensuing word because of their tendency to appear together. The classic Stroop task is an excellent paradigm to explore the cognitive mechanisms involved in conflict detection and resolution during lexical selection, given its simplicity in design. Moreover, the mechanisms elicited by the task are assumed to be domain general mechanisms, common to other cognitive tasks (see Larson, Clayson & Clawson, 2014, for a review of ERP findings on Flanker and Stroop), and directly related to selection in language use. By taking direction from the ERP and behavioral literature on paradigms involving lexical selection, it explores conflict-detection mechanisms in a paradigm that involves the selection of words within phrases. In particular, participants selected the target verb based on the presence of a noun based on familiar Verb-Noun multiword units. To investigate the neurophysiological correlates of selection, the distractors presented were manipulated across different trials. Through the comparison of trials in which only one verb is a plausible candidate (Example 1) to trials in which two verbs compete for selection (Example 2), this design allowed to explore the neural correlates of conflict detection during lexical selection in multiword units.

(1) eat – shoot breakfast familiarized target verb: EAT

(2) eat – skip breakfast

---

Part of the content of this chapter has been published as Pulido, Manuel F. & Paola E. Dussias, (2019). The neural correlates of conflict detection and resolution during multiword lexical selection: evidence from bilinguals and monolinguals. Brain Sciences, 9, 110. DOI: 10.3390/brainsci9050110 (Open Access)
The paradigm used here to examine lexical selection does not rely on the same cross-modal cues present in studies that have employed the Stroop paradigm (e.g., word-color) but instead relies on lexical cues; this allows us to investigate the same basic conflict-detection mechanisms widely investigated in the behavioral and ERP literature on semantic interference based on the Stroop effect.

Given previous findings of conflict-detection components in tasks that have been widely studied (including the Stroop and Flanker tasks), it is expected that a functionally similar component may be elicited when conflict arises during selection of words within multiword units. However, the time-course of conflict detection is expected to vary as a result of the demands of the task at hand. In the present study, rather than making selection contingent on modality (as in the classic Stroop), we designed a task in which the cue for selection is provided by a word that frequently co-occurs with the target. By making lexical choices dependent on the meaning of another word within a familiar multiword unit, the current study seeks to go beyond the single-word level. Here we explore conflict detection and resolution in a scenario in which candidates have already been activated; that is, the main manipulation in this paradigm lies on creating conditions that induce conflict in selection or not, by carefully controlling the lexical candidates presented (target and non-target verbs) before the cue for selection (i.e., the noun) appears. The issue of whether potential candidates are activated simultaneously or sequentially remains, to our knowledge, largely unexplored and goes beyond the scope of this paper (but see Dooghe, De Baene & Hartsuiker, 2016; Schriefers, 1993; Schriefers, Jescheniak & Hantsch, 2002, 2005) on a related topic concerning the activation of different determiners for a noun). Yet, our manipulation draws on the assumption that speakers activate potential candidates concurrently (Thierry & Wu, 2007; Oppenheim, Wu & Thierry, 2018; Wu, 2010). We are able to also control for what candidates are more strongly activated and in what order, by presenting two verbs sequentially prior to the display of the noun. To control what verb was processed first and second, the paradigm
manipulated the order of presentation of the Target (T) and Distractor (D) verbs. Half of the trials presented the target right before the noun (DT order), while the other half presented the verb to be discarded before the noun (TD). Table 3-1 presents examples of the 2x2 manipulation in an experimental trial (target: *eat – breakfast*); importantly, the distractor verbs are also potential candidates for selection, as they are the verbs of foil trials (e.g., *shoot photos*).

The experiment reported below will allow us to determine the time course and scalp topography of conflict detection during selection in English multiword units, in two different populations: native speakers of English and late Spanish-English bilinguals. While both groups completed the exact same task, the conflict-inducing distractors are potential candidates based on knowledge of English alone, but not based on knowledge of Spanish multiword units. This allowed us to explore conflict in lexical selection in the non-native language for Spanish-English bilinguals, and to consider the role of familiarity and entrenchment in the L2.

Table 3-1: Sample of experimental manipulation (*eat breakfast*).

<table>
<thead>
<tr>
<th>Distractor Plausibility</th>
<th>Order</th>
<th>Verbs</th>
<th>Noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implausible</td>
<td>DT</td>
<td>shoot – EAT&lt;sup&gt;9&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Implausible</td>
<td>TD</td>
<td>EAT – shoot</td>
<td>breakfast</td>
</tr>
<tr>
<td>Plausible</td>
<td>DT</td>
<td>skip – EAT</td>
<td></td>
</tr>
<tr>
<td>Plausible</td>
<td>TD</td>
<td>EAT – skip</td>
<td></td>
</tr>
</tbody>
</table>

<sup>8</sup> While we have devised a paradigm intended to create propitious conditions to elicit the type of conflict that we aim to study, we refrain from making any claims about the sequence followed during lexical access when speakers are forming or are retrieving multiword phrases. Indeed, in natural language the order of this process may be quite variable. For example, although a Verb-Noun sequence is a highly common order, other structures containing the same lexical combination reverse this order to Noun-Verb. To illustrate, structures such as relative clauses require speaker to produce a noun before the verb is selected (e.g., “they skipped the breakfast, which was in fact delicious” vs “the breakfast that they skipped was in fact delicious”). The latter might be more directly comparable with selection in our paradigm.

<sup>9</sup> Here and throughout the rest of the chapter, the target verb in each example is indicated by using capital letters. Verbs in the experiment were presented in lower case.
Materials and Methods

Participants

Fifty-five participants were recruited, of which thirty were native speakers of English and twenty-five were late Spanish-English bilinguals. All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Internal Review Board of the Pennsylvania State University (PRAMS0004598/00034810). Participants were paid 10 US dollars per hour of participation or per the corresponding fraction of an hour. All 23 native English speakers reported limited knowledge of a foreign language in a ten-point self-rated proficiency scale (mean: 3.37; SD: 2.01)\textsuperscript{10}. Four subjects were excluded because they reported being simultaneous bilingual speakers of English and another language. One Spanish-English speaker was excluded due to switched language dominance, as indicated by self-reported proficiency in each language (greater dominance in English). Five additional subjects were excluded because they did not complete the second session. The remaining participants included 23 native English speakers (17 female, 6 male) and 22 late bilinguals (12 female, 10 male). All participants were right-handed, reported normal or corrected-to-normal vision, and had no history of neurological disorders.

Materials

Three types of materials were created: (1) a list of 16 English multiword units (i.e., collocations) and their associated distractors; (2) materials for a familiarization phase; and (3) materials for a multiword-based lexical selection task. The latter two were based on the multiword units selected in (1). We first describe each in turn.

\textsuperscript{10}Self-reported proficiency measures were collected by administering the LEAP questionnaire (a language background questionnaire) completed by participants as part of the testing procedure described below.
**English multiword units (collocations)**

First, two lists of eight verb-noun target combinations (V-NP) were created. One list contained the target verb-noun multiword units (e.g., cancel [a] trip, eat breakfast), while the second list consisted of V-NP phrases used as foils (e.g., schedule [a] time; skip school). We will refer to these lists as targets and foils, respectively. All verb-noun units contained a noun that was congruent across the two languages (i.e., could be translated literally). To create the manipulation that examined the effect of conflict during lexical selection, two additional lists were of the following type: one list (the plausible distractor list) included combinations of foil-verb + target-noun combinations that were plausible (schedule [a] trip; skip breakfast). The final list (the implausible distractor list), included combinations of foil-verb + target-noun combinations that were implausible (break [a] trip; shoot breakfast). The four lists are given in Table 3-2.

In order to determine the plausibility and collocational status of the V-NP combinations in each of the four lists, two measures were obtained. First, plausibility ratings were collected on a scale from 1 (implausible) to 7 (plausible) through Amazon Mechanical Turk (MTurk). For all the participants, we collected language background information and excluded participants who reported not being native speakers of English. As an additional measure, we restricted IP addresses to the US. One participant who reported being a non-native speaker of English was excluded. Average plausibility ratings were calculated for responses from 29 native English speakers. The results showed that high plausibility ratings were given to target multiword units (e.g., eat breakfast; mean: 6.36; SD: 0.21), and foils (e.g., skip school; mean: 5.43; SD: 0.63), as well as to items in the plausible distractors list (e.g., skip breakfast; mean: 5.91; SD: 0.72). Implausible distractors were rated low in the plausibility scale (shoot breakfast; mean: 1.44; SD: 0.25). Based on Bonferroni-corrected pairwise comparisons, ratings for implausible non-targets were significantly lower both from those for targets ($t(39) = 72.86, p < .001$), foils ($t(39) = 45.49, p < .001$), and plausible
distractors ($t(39) = 35.7, p < .001$); also, in line with the corpus-based measures, ratings revealed that targets were preferred over (plausible) distractors ($t(39) = 3.89, p < .01$).

Table 3-2: List of Verb-Noun phrases per condition.

<table>
<thead>
<tr>
<th>Target list</th>
<th>Foil list</th>
<th>Plausible distractors</th>
<th>Implausible distractors</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancel [a] trip</td>
<td>schedule [a] time</td>
<td>schedule [a] trip</td>
<td>break [a] trip</td>
</tr>
<tr>
<td>eat breakfast</td>
<td>skip school</td>
<td>skip breakfast</td>
<td>shoot breakfast</td>
</tr>
<tr>
<td>finish [a] story</td>
<td>break [the] news</td>
<td>break [a] story</td>
<td>fail [a] story</td>
</tr>
<tr>
<td>promote peace</td>
<td>disturb [the] neighbors</td>
<td>disturb [the] peace</td>
<td>schedule peace</td>
</tr>
<tr>
<td>rent [a] movie</td>
<td>shoot photos</td>
<td>shoot [a] movie</td>
<td>launder [a] movie</td>
</tr>
<tr>
<td>save money</td>
<td>launder clothes</td>
<td>launder money</td>
<td>skip money</td>
</tr>
<tr>
<td>teach [a] class</td>
<td>fail [a] test</td>
<td>fail [a] class</td>
<td>cash [a] class</td>
</tr>
<tr>
<td>write [a] check</td>
<td>cash [a] prize</td>
<td>cash [a] check</td>
<td>disturb [a] check</td>
</tr>
</tbody>
</table>

To measure the collocational status in the V-NP sequences, t-scores (Evert, 2009; Gries, 2010, 2013) were calculated using data from the Corpus of Contemporary American English (COCA, Davies, 2008). T-scores have been shown to be more resistant to inflation than other association measures, such as Mutual Information (MI) scores, and allow for comparability with previous work (e.g., Wolter & Gyllstad, 2011). Word pairs with t-scores equal or greater than 2.0 are considered collocational (Hunston, 2002). The analyses revealed that all the target and foil items were collocational, i.e., conventional multiword units (mean: 26.82; SD: 17.73). T-scores were also calculated for the plausible and implausible distractors. All plausible distractors had t-scores of at least 5.0 (mean: 12.01; SD: 4.78), confirming their collocational status. Based on log-normalized frequencies from corpus data, target V-NP phrases were more frequent (mean: 6.3; SD: 1.15) than plausible distractors (mean: 5.05; SD: 0.76; $t(39) = 5.94, p < 0.0001$). Implausible distractors were in all cases non-collocational. While some observations of implausible distractors were expected due to chance, given the
large size of COCA (with over 570 million words), t-scores had in all cases negative values, well below the conventional threshold of 2.0 (min: -25.11; max: -1.44).

One final note is in order. Because the same materials were used with both groups of speakers (i.e., native and non-native), the proportion of cognates between English and Spanish was controlled, with 50% of cognate nouns in the target and foil lists. The experimental manipulation for the bilingual group also required that items in the target list be congruent between the two languages, whereas the verbs in plausible distractors needed to be incongruent between the two languages. Language congruency was determined by the two authors, who are native speakers of Spanish. Additionally, verbs were normed by collecting translation data from a group of 22 Spanish-English bilinguals (mean self-rated English proficiency on a 10-point scale: 8.75; SD: 0.78). Bilinguals were asked to translate the isolated verb of each English collocation (e.g., *eat, skip*) into Spanish, and the verbs of the Spanish equivalents into English (e.g., *comer, saltarse*). The combined proportion of target responses was calculated for each collocation. Target responses for the translations of the verbs of congruent collocations averaged 94% accuracy; for incongruent collocations, accuracy was only 38%.

**Familiarization phase**

A prior familiarization phase was completed so that the target verb could be correctly identified. For each item in the target list, a carrier sentence was created that contained the target V-NP combination. All sentences contained the target V-NP phrase (e.g., “eat breakfast”), which was located towards the middle of the sentence. Participants read sentences at their own pace.

(2) Carrier sentence: “On Sundays I usually get up later and eat breakfast in my pajamas”

Target multiword unit: *eat breakfast*
To ensure that the correct target V-NP units could be remembered, participants were given a gap-filler task in which they identified the correct multiword units in the sentences they had just read. The same sentences were presented again, which this time contained a gap instead of the target multiword unit. Two options were presented on the screen below the sentence, on the left and right side of the screen, as illustrated in Example 3.

(3) On Sundays I usually get up later and ____________ in my pajamas.

   eat breakfast  skip school

For each item, the non-target phrase was the associated foil presented in Table 3-2. The correct option (left or right) was selected by pressing the ipsilateral button on a Chronos button box; sentences were presented using E-Prime 3.0 (Psychology Software Tools, Pittsburgh, PA.). Participants were advised to pay close attention to the options presented as they would need to remember them in the ensuing task.

**Multiword-based Lexical Selection Task**

As described above, each multiword unit (e.g., *eat breakfast*) was matched with a plausible distractor (*skip breakfast*) and with an implausible distractor (*shoot breakfast*). In this task, participants were presented with the two verbs (target and distractor, e.g., *EAT – skip*) followed by the noun (*breakfast*). Once the noun was presented, participants were able to select the appropriate verb, based on the phrases presented during the Familiarization task (*eat* in this example). The order of presentation of target and distractor was counterbalanced. That is, *EAT – skip – breakfast* appeared for half of the trials in which *eat* was the target verb, and *skip – EAT – breakfast* in the other half. To control for the order of processing and activation of each verb (target and distractor), the two verbs were presented sequentially. At the beginning of each trial, one fixation cross appeared at the center of the screen for 200 ms. One verb then appeared to the left of the fixation cross; 200 ms later, a second verb appeared to the right of the fixation cross; both were displayed together for 300 ms, after which the verb on the left disappeared while the verb on the right remained on the screen for an
additional 200 ms. The noun was then displayed for 300 ms, followed by a fixation cross that remained on the screen until a response was made. Responses were made by pressing the right or left button on the button box, to select the verb that was presented on the corresponding side of the screen. To ensure that participants were able select the target verbs, the correct verb (e.g., *eat*) was then presented as feedback for 400 ms (in blue font for correct responses, and in red font for incorrect responses). Words were displayed in font Arial size 30. The task was presented using E-Prime 3.0. A sample trial can be seen in Figure 3-1.

Figure 3-1: Sample trial of the Lexical Selection Task.

The manipulation of Plausibility of the distractor (plausible vs implausible), altogether with the Order of presentation of Target and Distractor verbs (T-D vs D-T), produced a 2x2 design. To generate a sufficient number of trials for the ERP analysis, each of the eight multiword units in each list was presented five times in each of the four conditions, producing 40 trials per condition (for a total of 320 trials). An additional ten practice trials were presented at the beginning of the task. Completing of the task took approximately 20 minutes.
Procedure

The study was divided into two separate sessions. In session 1, participants were consented after arriving in the lab. They then completed an English verbal fluency task to measure vocabulary knowledge and lexical access in production. In the verbal fluency task, participants are given 30 seconds to name exemplars for each of four categories presented (two animate); the score is calculated as the sum of total valid responses (i.e., no repetitions) generated for each category in the 30 seconds allotted. General English proficiency was then measured through an abridged version of the Michigan English Language Institute College English Test (MELICET, 2001) administered on a computer through Qualtrics (Qualtrics, Provo, UT), which contained a combination of 50 grammar and cloze-test items to assess overall language proficiency. Two behavioral measures of cognitive control were collected: the Stroop task and the Flanker task. Finally, to assess participants’ linguistic background, they completed the LEAP-Q (Marian, Blumenfeld & Kaushanskaya, 2007).

Participants came back to the lab for the second session and were consented again. In this session they completed the Familiarization, followed by the Multiword-based Lexical Selection Task. Each session took approximately 1h and 30 min.

The procedure for both groups (native and non-native speakers) was identical with the exception that the non-native speakers completed an additional task. Because the verbs of foil trials (which were also used as plausible distractors) were incongruent with Spanish and specific to English, participants were administered a Familiarity Rating task at the end of the experiment. A list with all the targets and foils was presented, for which participants were asked to rate the familiarity of each Verb-Noun Phrase multiword unit along a scale from 1 (completely unfamiliar) to 7 (completely familiar).
**EEG recording and analysis**

The continuous electroencephalogram (EEG) was recorded from 32 electrodes mounted in an elastic cap (EasyCap; Brain Products, GmbH) and an ActiChamp amplifier (Brain Products, GmbH) with a 24-bit analog to digital conversion (online sampling rate: 500 Hz, 0.1-100 Hz band-pass filter). Electrode impedances were kept below 5 KΩ. During recording, electrodes were referenced to the right mastoid. Grounding electrodes were mounted on the forehead and beneath the right eye. Blinks and eye-movements were measured by placing bipolar pairs of vertical (VEOG) above and below the left eye and lateral (HEOG) electrodes at the outer canthi of both eyes. Preprocessing steps and analyses were performed with MATLAB (R2016a, The Matworks, Inc.) and a combination of scripts and routines implemented in EEGLAB (v. 13.5.4b, Delorme & Makeig, 2004) and ERPLAB (v. 5.0, Lopez-Calderon & Luck, 2014). Datasets were filtered offline with a 25 Hz low pass and 0.1 Hz high pass noncausal IIR Butterworth digital filter (Tanner, Morgan-Short & Luck, 2015). Segments with excessive muscular artifacts on the continuous data were manually rejected. Subsequently, an independent component analysis (ICA) was performed to extract and reject remaining ocular and muscular artifacts, following Jung et al. (2000). As indicated below, separate ERP analyses were conducted for the groups of monolingual and bilingual speakers. The average number of independent components rejected during ICA averaged 3.3 (max: 5) in the monolingual dataset and 2.26 (max: 4) in bilinguals. Epochs ranging from -200 to 800 ms after onset of the noun were extracted from the pre-processed data. All epochs with activity exceeding ±100 μV at any electrode site were automatically removed using peak-to-peak moving window. Less than 0.1% of data were lost due to artifact rejection in each group. Baseline correction was done relative to pre-stimulus activity. Inaccurate trials were excluded from the analysis (8.2% in monolinguals, 7% in bilinguals). Data from one bilingual participant was excluded due to excessive artifacts.

Based on the predictions, described in the next section, our analysis of the EEG data focused on two main time windows. In order to analyze differences in processing of the noun,
we analyzed the canonical N400 window (300-500 ms post-stimulus presentation). Secondly, we predicted that conflict would emerge in trials in which a competitor had to be suppressed; the N_{inc} conflict-detection component was predicted to emerge at some point after processing of the noun and before a manual response was made. The analysis of the behavioral data determined that button presses started shortly after 800 ms, restricting the time-window of potential interest to 500-800 ms. In order to objectively determine the time window of conflict detection, running t-tests were performed using the FDR-corrected mass univariate test of the Factorial Mass Univariate Toolbox (FMUT; Fields, 2017), including Order and Plausibility as within-subject factors. The raw data was averaged into bins of 4 ms, and analyzed from 500 to 800 ms post-stimulus. While this type of test is quite stringent and likely to underestimate a significant effect on the data, it allows to examine the time-course and scalp distribution of the effect in a more exploratory fashion. All scalp sites were considered of interest with the exception of Fp1, Fp2, T7, T8, Oz, O1, O2, which were excluded. The significant comparisons resulting from running t-tests were then used to determine the time window for the analysis with repeated-measures ANOVAs; these are reported in the Results section and in Figures 3-4a-b. The ANOVAs were conducted with mean amplitudes as dependent variables, and distractor Plausibility and Order as independent variables. Midline ANOVAs included Frontality as a predictor (frontal, central, parietal) and were performed on Fz, Cz and Pz. Lateralized ANOVAs with Frontality and Hemisphere (left, right) predictors were conducted on F3, F7, F8, FC1, FC2, FC5, FC6, C3, C4, CP1, CP2, CP5, CP6, P3, P4, P7, P8. Greenhouse-Geisser corrected values are reported where appropriate. In order to better characterize the topography of significant effects, pairwise t-tests were performed on each scalp region (right frontal: F4, F8, FC2, FC6; left frontal: F3, F7, FC1, FC5; right posterior: CP2, CP6, P4, P8; left posterior: CP1, CP5, P3, P7). Results of post-hoc paired pairwise t-tests are reported with FDR corrected p values.

Finally, we performed correlation tests to investigate a potential association between the amplitude of individual subjects’ ERP components and performance in lexical selection as measured by RTs and accuracy. Individual ERP measures were calculated as the difference
waves of average amplitudes per condition (Plausibility, Order); behavioral scores were calculated as the difference between averages for each condition (e.g., RTs of Plausible DT – RTs of Implausible DT). Correlation tests were conducted between subjects’ individual ERP measures and RTs per condition, as well as between ERPs and accuracy.

**Predictions**

We predicted three different types of costs. One first prediction is concerned with processing of the noun (300-500 ms window). It is now well established by a large number of studies that semantic access is indexed by the N400 (300-500 ms window). Based on previous behavioral (Arnon & Snider, 2010; Durrant & Doherty, 2010; Wolter & Gyllstad, 2011) and EEG studies on processing of multiword units (Molinaro & Carreiras, 2010; Siyanova-Chanturia, Conklin, Caffarra, Kaan & van Heuven, 2017) we predict that nouns immediately preceded by the target verb (e.g., breakfast in eat breakfast) or by a plausible distractor (breakfast in skip breakfast) would be primed (manifested in the form of an attenuated N400); however, nouns preceded by an implausible distractor (e.g., breakfast in shoot breakfast) would show a cost in processing (i.e., an enhanced N400). Our second prediction is related to conflict detection in lexical selection. A conflict-detection component (i.e., the N_{inc}) will be elicited by trials in which distractors are plausible candidates for selection (e.g., skip in skip-EAT breakfast). Our prediction regarding the timing of conflict detection is that the N_{inc} component will be elicited at some point between processing of the noun and the average response time for the selection of the verb. Consistent with the literature reviewed above, we expect this component to arise over left centroparietal electrodes. Our third prediction considers the order of the verbs (target last vs distractor last), which is expected to affect the selection of the target. Based on a recency effect, we predict an asymmetry in cost affected by the order of presentation of target and distractor. Specifically, a greater cost in TD trials is predicted due to the need to discard the most recently (and therefore more strongly) activated
candidate. We predict that the distractor-last order might in fact modulate inhibition-\( \text{e.g., } N_{\text{inc}} \) or other selection-related components.

A final, more exploratory prediction is that components indexing conflict detection during lexical selection might be correlated with behavioral responses. Although previous studies found no correlation between the peak latency of the \( N_{\text{inc}} \) and RTs, we hypothesized that the extent to which conflict-detection components are present might be a predictor of behavioral performance. Therefore, in the results section below we present an exploratory analysis to investigate a potential correlation between the amplitude of individual subjects’ ERP components and performance in lexical selection measured by RTs and accuracy.

**Predictions for non-native speakers**

While the same predictions apply to both native and non-native speakers, there were some additional considerations concerning lexical selection in non-native speakers. First, as mentioned, potentially slower processing in L2 speakers was expected to be reflected in slower RTs as well as in delayed ERP components, relative to native speakers. However, the more critical prediction concerned the degree of interference that would be experienced by non-native speakers in trials in which the distractors presented were plausible, but only to the extent that speakers had experience of the multiword units in their L2. More specifically, we predict that the degree of interference for non-native speakers will be somewhat variable, and contingent on knowledge and familiarity with multiword units specific to English. Based on the assumption that the amount of interference would be reflected in the amplitude of conflict-related components, we predict a potential correlation between familiarity with the L2-specific multiword units and ERP difference waves.
Behavioral results

This section first presents the results of the individual differences measures collected during the first session of the study, and the results of the Familiarization phase. The following subsections report on the analysis of behavioral and ERP data from the multiword-based Lexical Selection task.

The individual measures collected in session 1 served the double purpose of allowing us to characterize the tested population in detail and providing individual measures to be included in subsequent analyses. Table 3-3 below presents the results of the English verbal fluency task, Stroop and Flanker effects, the MELICET test, as well as linguistic background information collected through the LEAP-Q. The English verbal fluency score is the total number of valid responses. The Flanker effect was calculated by subtracting the RTs of congruent trials (all arrows pointing in the same direction) from incongruent trials (target and distractor arrows pointing in opposite directions). To obtain the Stroop effect, the RTs of baseline trials (naming the font color of a sequence of several @s) were subtracted from incongruent trials (naming the font color of, e.g., the word “BLUE” presented in red font). Stroop results are excluded for one participant due to technical failure.

Table 3-3 also presents the results of the multiword unit familiarity rating task completed by bilinguals. The data show that non-native speakers were familiar with the expressions, but also that familiarity was lower and more variable for L1-L2 incongruent expressions, as reflected by the larger SD.

To ensure that participants had completed the Familiarization phase adequately, we examined the accuracy of responses collected on the gap-filler task (see example 3 above), which was 96% (min: 83.33, SD: 5.42) for native speakers and 93.24% (min: 83.33%, SD: 4.72) for non-native speakers. This confirmed that participants in both groups could correctly select the conventional multiword units in context. In what follows, we first report the behavioral results of the experimental trials in the Multiword-based Lexical Selection task, and then the ERP analysis.
Table 3-3: Summary of cognitive and proficiency measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Native</th>
<th>Non-native</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (in years)</td>
<td>23.09 ± 3.6</td>
<td>32.17 ± 5.97</td>
</tr>
<tr>
<td>L1 self-rated proficiency</td>
<td>9.83 ± 0.36</td>
<td>9.84 ± 0.36</td>
</tr>
<tr>
<td>L2 self-rated proficiency</td>
<td>3.37 ± 2.01</td>
<td>8.58 ± 0.63</td>
</tr>
<tr>
<td>MELICET (/100)</td>
<td>90.7 ± 5.07</td>
<td>78.27 ± 11.8</td>
</tr>
<tr>
<td>English Verbal Fluency</td>
<td>52.48 ± 11.08</td>
<td>42.14 ± 8.41</td>
</tr>
<tr>
<td>Flanker effect (ms)</td>
<td>48.87 ± 23.65</td>
<td>47.85 ± 23.64</td>
</tr>
<tr>
<td>Stroop effect (ms)</td>
<td>182.2 ± 100.06</td>
<td>125.81 ± 153.79</td>
</tr>
<tr>
<td>Familiarity L1-L2 congruent MWUs (/7)</td>
<td>6.45 ± 0.65</td>
<td></td>
</tr>
<tr>
<td>Familiarity L1-L2 incongruent MWUs (/7)</td>
<td>5.51 ± 1.23</td>
<td></td>
</tr>
</tbody>
</table>

**Accuracy in the Lexical Selection task**

The proportion of correct responses was calculated for each of the experimental conditions in the 2x2 design. To recapitulate, the manipulations involved Group (Native, Non-Native), Plausibility (plausible vs implausible distractors), and Order (Distractor and Target vs. Target and Distractor; DT vs TD). Table 3-4 presents the accuracy results by condition for both groups of speakers.

The effect of Plausibility and Order was examined in both groups by performing a mixed-effect logistic regression. This analysis is ideal for binary variables such as accuracy, as it allows to analyze the unaggregated data rather than means (Jaeger, 2008). The analyses of the behavioral data reported here and in other sections were carried out with the lme4 package (Bates, Maechler, Bolker & Walker, 2015) in R version 3.3.2 (R Core Team, 2016).

Following attempts to build maximally specified models (Barr, Levy, Scheepers & Tily, 2013), which led to convergence issues, the random effects structure was simplified.
(Bates, Kliegl, Vasishth & Baayen, 2015). Final models included random intercepts for subjects and items (Baayen, Davidson & Bates, 2008). To control for variability in the plausibility of verbs, models also included by-subject slopes for the average normed plausibility of the distractors.

Table 3-4: Results of accuracy in selection of the target verb by condition.

<table>
<thead>
<tr>
<th>Distractor Plausibility</th>
<th>Order</th>
<th>Mean Acc. %</th>
<th>SD</th>
<th>Mean Acc. %</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Native</td>
<td>Non-native</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implausible</td>
<td>DT</td>
<td>98.37</td>
<td>12.67</td>
<td>96.02</td>
<td>19.55</td>
</tr>
<tr>
<td></td>
<td>TD</td>
<td>96.96</td>
<td>17.19</td>
<td>97.39</td>
<td>15.96</td>
</tr>
<tr>
<td>Plausible</td>
<td>DT</td>
<td>87.5</td>
<td>33.09</td>
<td>90.57</td>
<td>29.24</td>
</tr>
<tr>
<td></td>
<td>TD</td>
<td>84.57</td>
<td>36.15</td>
<td>88.41</td>
<td>32.03</td>
</tr>
</tbody>
</table>

Fixed effect factors included Group, Plausibility and Order, and their interaction, to analyze the effect of the experimental manipulation. The contribution of cognitive control was investigated by including fixed effects for the Stroop and Flanker Effect. English verbal fluency scores were also included as an index of lexical knowledge. Age was also included in order to control for potential differences across groups. All continuous variables were centered.

Model selection started with the mixed-effects model described. In a step-by-step backward model selection procedure, predictors were removed one by one, but were kept if the model fit was significantly improved by including the predictor (likelihood ratio test, $p < .05$). Parameter-specific $p$ values were estimated using the normal approximation. Because Stroop results from one participant were not available, data from that subject were excluded in the initial analysis. However, the final model selected did not include Stroop, Flanker or English Verbal Fluency scores, as they did not significantly improve the fit of the model. The analysis was run again on the full dataset with the additional participant, and significance
results and parameter estimates did not substantially differ. For completeness, we report the results of the parameters in the analysis that was performed on the full dataset.

The reference levels were set to native speaker for Group, implausible distractor for Plausibility, and DT for Order. The results revealed a main effect of Group ($\beta$: -0.96, SE: 0.34, $p < .05$), showing that non-natives were significantly less accurate than native speakers. There were also main effects of Plausibility ($\beta$: -2.37, SE: 0.31, $p < .0001$), with a lower accuracy when distractors were plausible relative to implausible; and Order, with more errors being produced in Target-Distractor trials ($\beta$: -0.65, SE: 0.32, $p < .05$). Additionally, two- and three-way interactions arose for Group x Plausibility ($\beta$: 1.31, SE: 0.38, $p < .001$), Group x Order ($\beta$: 1.01, SE: 0.42, $p < .01$), and Group x Plausibility x Order ($\beta$: -1.14, SE: 0.47, $p < .05$). These interactions reveal that, in bilingual speakers, implausible distractors in TD order had the opposite effect than in natives, resulting in higher accuracy (as shown by the positive $\beta$ estimates). This is not the case in trials with plausible distractors, which have the same effect in both groups, as shown by the negative coefficient of the three-way interaction. While the interactions are somewhat complex, the pattern is clearly illustrated in Figure 3-2 below.

The full model output is available in Appendix D (Table D1).

Figure 3-2: Accuracy in selection of the target verb by condition. Error bars indicate 95% confidence intervals.
**Reaction Times in the Lexical Selection task**

In order to prepare the data for RT analysis, inaccurate responses were first removed. Trials with RTs outside the absolute thresholds of 300 ms – 4000 ms were also excluded from the analysis. RTs were z-score normalized, and outliers were then removed for each participant based on their individual median absolute deviation (MAD). Using the MAD method is a more robust measure for outlier removal than standard deviations, given that the latter are susceptible to be distorted by observations that strongly deviate from the mean (Levshina, 2015, pp. 48-50). Trials above or below 3 absolute deviations from the median were excluded. Table 3-5 presents a summary of the mean RTs by condition for each group of speakers.

Table 3-5: Reaction Times in selection of the target verb by condition.

<table>
<thead>
<tr>
<th>Plausibility</th>
<th>Order</th>
<th>Native Mean RT</th>
<th>Native SD</th>
<th>Non-native Mean RT</th>
<th>Non-native SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implausible</td>
<td>DT</td>
<td>815.82</td>
<td>348.02</td>
<td>977.31</td>
<td>363.71</td>
</tr>
<tr>
<td>Implausible</td>
<td>TD</td>
<td>838.69</td>
<td>337.65</td>
<td>994.7</td>
<td>349.05</td>
</tr>
<tr>
<td>Plausible</td>
<td>DT</td>
<td>858.94</td>
<td>399.93</td>
<td>1027.23</td>
<td>413.35</td>
</tr>
<tr>
<td>Plausible</td>
<td>TD</td>
<td>889.91</td>
<td>385.65</td>
<td>1043.07</td>
<td>394.62</td>
</tr>
</tbody>
</table>

A linear mixed-effects regression analysis was performed on the normalized RTs. Models included the same mixed-effects structure described for the accuracy data, and were built and selected using the same procedures as in the logistic regression models. The model selected contained fixed-effects for Group, Plausibility and Order. Their interactions as well as the other behavioral measures of individual differences considered (English Verbal Fluency, Stroop and Flanker) did not improve the model fit and were not part of the final model.

The main effect of Group confirmed that Spanish-English bilingual speakers were significantly slower than native English speakers ($\beta$: 0.39, SE: 0.16, $p < .05$). In line with the
accuracy results, the analysis revealed a main effect of Plausibility, resulting in longer RTs in trials with plausible distractors ($\beta$: 0.13, SE: 0.02, $p < .0001$). Likewise, the Order of Target-Distractor trials caused a significant delay in RTs relative to Distractor-Target trials ($\beta$: 0.06, SE: 0.02, $p < .01$). The full model output is also available in Appendix D (Table D2). Figure 3-3 illustrates the effect of Order and Plausibility by Group. To examine whether the cost caused by the main effects was cumulative, we conducted a follow-up comparison across the four conditions with FDR-corrected pairwise t-tests, separately for each language group. For both groups, the results revealed significant differences between plausible TD and both implausible DT trials (all $p < .001$) and implausible DT (all $p < .05$). Differences between implausible DT and plausible DT were significant in non-native speakers ($p < .05$) and marginally significant for monolinguals ($p = .05$).

**Figure 3-3:** RTs for selection of the target verb by Group. The figure shows the RTs of trials with correct responses. Error bars indicate 95% confidence intervals.

**Discussion of behavioral results**

Implausible DT trials (e.g., *shoot – EAT – breakfast*) represent the baseline level, both conceptually and in the statistical analyses. The implausibility of the distractor verb-
noun combination means that the level of conflict between distractor and target verbs is kept to a minimum. Likewise, the fact that the target verb follows the distractor and appears just before the noun, should produce no cost. If anything, it should produce facilitation in the form of “collocational priming” (Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita; 2015; Yamashita & Jiang, 2010), given that the target verb (eat) is followed by a predictable noun (breakfast), in consonance with previously experienced multiword units. The current paradigm reveals the expected behavioral costs from two independent factors. The results showed that the simple effect of Order (as seen in the implausible distractor with TD order condition) caused a significant slowdown in RTs. Distractor Plausibility had a similar effect, as seen in plausible DT trials, in which the cost is generated by a plausible distractor, but with no additional cost on account of the baseline DT order. The reported pairwise comparison confirmed that differences in significance increased as effects accumulate. Specifically, plausible TD trials, which induced a cost from both factors, show that the cost is in fact doubled, suggesting a cumulative effect of selecting between two plausible candidates when the distractor also precedes the noun.

However, the results of the accuracy analysis are only partly congruent with the RT data, and provide additional insight into the nature of the costs of selection across groups. Accuracy in plausible-distractor trials largely mirrored the effects of Order and Plausibility across both groups. On the other hand, the interactions of these two factors with Group revealed a different picture when implausible distractors had to be rejected. Native speakers’ accuracy was lower when an implausible distractor immediately preceded the noun, indicating that processing of an implausible V-N sequence (e.g., shoot breakfast) incurred a cost. On the contrary, this resulted in an advantage for L2 speakers, suggesting that processing was not linear, unlike that of natives. That is, the advantage in rejecting implausible phrases seems to indicate that Verb and Noun were not processed as a unit, but rather that bilingual speakers might have focused on the plausibility of each candidate. This interpretation would be in line with theories that propose that non-native speakers’ processing relies more on plausibility and less on other linguistic cues than native speakers’ (e.g., the
Shallow Structure Hypothesis, Clahsen & Felser, 2006). We return to this idea and discuss it in more depth in the general discussion.

**ERP results**

We present the results of the ERP analysis in two subsections, one for each of the language groups. Given that there is no prior literature that has identified the ERP components of conflict detection and resolution during multiword lexical selection, we take an exploratory approach to define the time window for analysis that is relevant for each group. To do this, we first relied on mass univariate analysis to identify the time windows of interest for the analysis, and then conducted repeated measures ANOVAs on those time windows. This approach involves conducting a statistical test at each time point and electrode of interest that is then corrected to control the Type I error or false discovery rate, offering a more objective and systematic way to determine time windows and regions of interest (see the section “EEG recording and analysis” for details of implementation in this analysis).

Additionally, because there was no a priori justification to assume identical ERP components in native and non-native speakers, we conducted separate analyses for each group. This decision is grounded on a considerable number of processing studies showing that processing in L1 and L2 speakers may elicit different ERP components and/or in different time windows, even when conducting the same task with identical stimuli (Morgan-Short, 2014; Steinhauer, White & Drury, 2009; Tanner, McLaughlin, Herschensohn & Osterhout, 2013).

**ERP analysis of native English speakers**

The ERP analysis focused on three different time windows: the 300-500 ms time window (the N400 component), and two time windows (500-600 ms and 650-800 ms) revealed by significant differences in the exploratory mass univariate analysis. This analysis
revealed significant differences modulated by Order, mostly distributed over right frontal and left posterior electrodes (F4, F8, FC6, FC2, CZ, C4, CP1, CP5, C4, CP2, PZ, P3). Significant differences between waves arose between 500 and 600 ms after presentation of the noun, as well as at a later time interval around 730 ms. Significant time windows are illustrated by grey bars in Figures 3-4a-b for two representative electrodes (plots for six additional representative electrodes are provided in the supplementary materials). Because mass univariate analysis is likely to underestimate the true boundaries and differences between waveforms due to stringent correction procedures (Fields, 2017; Groppe, Urbach & Kutas, 2011) we conducted additional ANOVAs in the time windows around these peaks, between 500-600 ms and 650-800 ms. The results for these three time windows analyzed are presented below, and scalp maps are presented in Figure 3-5.

**N400 component (300-500 ms)**

The midline ANOVA revealed a main effect of Order ($F(1, 22) = 11.57, p < .05$), showing a significant difference in the N400 of DT and TD trials. There was also a main effect of Frontality ($F(1.07, 23.58) = 32.17$), but no interactions and no effect of Plausibility. The lateralized ANOVA also revealed main effects of Order ($F(1, 22) = 12.54$), and Frontality ($F(1.07, 23.5) = 50.78$). Additionally, there were three-way interactions between Plausibility, Hemisphere and Frontality ($F(2, 44) = 4.79$), and Order, Hemisphere and Frontality, ($F(2, 44) = 9.82$), showing that the effect of both Plausibility and Order differed across scalp regions. In post-hoc pairwise comparisons, we examined the simultaneous effect of Plausibility and Order for each scalp region. Across all regions, there was a significant effect of Order for trials with implausible distractors; that is, TD trials (e.g., *EAT* – *shoot* – *breakfast*) elicited a greater negativity than DT trials (*shoot* – *EAT* – *breakfast*). But when both verbs were plausible (e.g., *EAT* – *skip* – *breakfast*), differences elicited by Order were only significant in right frontal electrodes, with TD eliciting a greater N400 relative to DT ($t(91) = -4.48, p < .001$).
Additionally, in trials in which distractors preceded the noun (TD), the effect of Plausibility (e.g., \textit{EAT} – \textit{skip} – \textit{breakfast} vs \textit{EAT} – \textit{shoot} – \textit{breakfast}) was most significant over left posterior electrodes ($t(91) = 4.87, p < .0001$), and only marginally significant in left frontal electrodes ($t(91) = 2.14, p = .07$). In Distractor-Target trials, Plausibility was not significant over any of the regions. Overall, differences were most prominent in the right frontal region, where comparisons were significant across all conditions except for implausible and plausible Distractor-Target trials. In left and right posterior electrodes, differences reached significance between implausible TD trials and each of the other conditions only. Figures 3-4a-b show two representative electrodes.

\textbf{500-600 ms window}

The analysis of the midline electrodes showed main effects of Order ($F(1, 22) = 5.24$, $p < .05$) and Frontality ($F(1.09, 24.02) = 25.21, p < .0001$). Similarly, the lateralized ANOVAs also revealed main effects of Order ($F(1, 22) = 7.22, p < .05$) and Frontality ($F(2,44) = 31.67, p < .0001$), as well as three-way interactions of Plausibility, Hemisphere and Frontality ($F (2, 44) = 5.5, p < .01$), and Order, Hemisphere and Frontality ($F (2,44) = 16.08, p < .0001$). Post-hoc comparisons revealed significant differences between implausible Distractor-Target trials (\textit{shoot} – \textit{EAT} – \textit{breakfast}) and all other trial types in left frontal, and left and right posterior electrodes; differences were most highly significant in the left posterior region ($p < .0001$). This confirmed that the negative peak between 500-600 ms was present in all conditions except for implausible Distractor-Target trials. Over right frontal electrodes, plausible Distractor-Target (\textit{skip} – \textit{EAT} – \textit{breakfast}) was also different from both implausible TD (\textit{EAT} – \textit{shoot} – \textit{breakfast}; $t(91) = 4.84, p < .0001$) and plausible TD (\textit{EAT} – \textit{skip} – \textit{breakfast}; $t(91) = -2.61, p < .05$).
650-800 ms window

As in the other two time windows analyzed, main effects of Order \( (F(1, 22) = 7.54, p < .05) \) and Frontality \( (F(0.9, 19.84) = 14.26, p < .001) \) were revealed by the midline electrode analysis. In the lateralized ANOVA, the same effects were found for Order \( (F(1, 22) = 7.28, p < .05) \) and Frontality \( (F(1.1, 23.28) = 17.5, p < .0001) \), and again, three-way interactions between Plausibility, Hemisphere and Frontality \( (F(2, 44) = 5.28, p < .01) \) and Order, Hemisphere and Frontality \( (F(2, 44) = 25.39, p < .0001) \). Pairwise t-tests revealed no differences in any of the comparisons in left frontal and right posterior regions. In right frontal and left posterior electrodes, differences were significant for each comparison, with the exception of implausible vs plausible Target-Distractor trials \( (EAT – shoot – breakfast vs EAT – skip – breakfast) \); right frontal, \( t(91) = -0.79, p = .44 \); left posterior, \( t(91) = -0.47, p = .64 \).
Figure 3-4: ERP waveforms for native speakers at (a) FC6 (right frontocentral) and (b) CP1 (left centroparietal). The legend shows examples for each condition, with the target verb capitalized. Grey bars underneath show the time points of significant effects in running t-tests within the pre-defined 500-800 ms time window. Boxes show the three time windows analyzed. The arrow in (b) indicates the Ninc effect. Negativity is plotted up.

Figure 3-5: Scalp topographies of the difference waves showing the effects of (a-b) Distractor Plausibility (Plausible – Implausible) and (c-d) Order (TD – DT in plausible trials) in the 500-600 and 650-800 ms time windows, respectively.
Correlation between neurophysiological and behavioral data in native speakers

Finally, exploratory tests were performed to investigate the potential correlation between ERP amplitude difference waves and individual behavioral data. To investigate the correlation between the effect of Plausibility in ERP and behavioral data, first the individual \( N_{inc} \) amplitudes were calculated by subtracting the average amplitudes of Plausible DT-Implausible DT in the 500-600 window. The same subtractions were performed on the individual averages of RT and accuracy data. The results revealed a marginally significant correlation between the \( N_{inc} \) and the difference in RTs of Plausible DT-Implausible DT (\( r = - .38, p = .08 \)). However, no correlation was found between individual \( N_{inc} \) measures and accuracy.

To investigate the effect of Order, the ERP difference wave was calculated for TD-DT comparison within the 650-800 ms time window for both the Plausible and Implausible distractor conditions over left posterior and right frontal electrodes. The same subtractions were performed on individual RT and accuracy averages. In contrast with the previous comparison, this time the results revealed a significant correlation between the ERPs and the proportions of accurate responses between Plausible TD-Plausible DT in both left posterior (\( r = -.48, p < .05 \)) and right frontal (\( r = -.54, p < .01 \)). However, no correlation was found between individual ERP measures from 650-800 ms and RTs. For Implausible trials, correlations were non-significant. We further interpret these results in the Discussion.

Discussion of ERP results in native speakers

In the analysis of the monolingual ERP data we aimed to investigate the neural correlates of conflict detection and resolution in native speakers of English, in a task in which they had to select the correct verb out of two options presented, to form a familiar multiword unit. Differences across conditions were examined by comparing trials in which the distractor verb was plausible or implausible, and based on order (target verb presented last or distractor
presented last). In the ERP analysis we further disentangle different types of cost by examining three time windows locked to the presentation of the noun. The analysis of the N400 window allowed us to examine differences in processing of the noun resulting from priming, or lack thereof, from verbs that preceded it. The results showed the expected modulations within the prototypical time window and parietal scalp distribution of the N400. Left and right posterior electrodes reflected significant differences between conditions in which the noun was primed by the verb relative to the one condition in which the verb was an implausible distractor. That is, in Distractor-Target trials, in which the noun was immediately preceded by the target verb (both \textit{skip} – \textit{EAT} – \textit{breakfast} and \textit{shoot} – \textit{EAT} – \textit{breakfast}), the N400 was attenuated, regardless of distractor plausibility. Critically, the same was true for trials in which the noun was preceded by a distractor that was plausible (\textit{EAT} – \textit{skip} – \textit{breakfast}). However, as illustrated in Figure 3-4b, these conditions were all different from cases in which there was no priming possible, namely, in Target-Distractor trials that presented an implausible verb before the noun (\textit{EAT} – \textit{shoot} – \textit{breakfast}). Interestingly, right frontal electrodes (as illustrated in Figure 3-4a) revealed a gradient effect modulated by both Plausibility and Order, such that nouns primed by both the target and a plausible distractor show the most attenuated N400.

The following two time windows based on the mass univariate analysis (500-600 ms; 650-800 ms) allowed us to address the critical predictions regarding the selection of the verb in native English speakers. In the 500-600 ms time window, a negative-going peak was elicited in the two conditions in which conflict was caused by a competing plausible verb (black lines in Figure 3-4). The scalp topography of this effect (Figure 3-5a) was also consistent with the left centro-parietal distribution of the N\textsubscript{inc} effect reported in previous studies (Appelbaum, Meyerhoff & Woldorff, 2009; Dyer, 1971; Glaser & Glaser, 1982; Larson, Kaufman & Perlstein, 2009; Liotti, Woldorff, Perez & Mayberg, 2000; West, 2003; West & Alain, 1999). The result of the N\textsubscript{inc} elicited by plausibility-induced conflict is consistent with our second prediction regarding conflict due to interference in selection. Namely, we expected that the N\textsubscript{inc} would be elicited by trials in which distractors are plausible
candidates for selection (e.g., *skip in skip – EAT – breakfast*), at some point between processing of the noun and before responses were made.

Additionally, the same negative deflection appeared to be elicited by the Order effect alone within the 500-600 ms window, in trials with an implausible distractor in which the non-target verb was presented last (i.e., implausible TD, *EAT – shoot – breakfast*). This would suggest a cost of Order; however, it must be noted that the comparison between plausible TD and plausible DT for that same time window produced no effect of Order. Because the implausible TD trials are the only ones in which a greater negativity was elicited during the N400 time window, it seems possible that this sustained negativity may be a carryover effect. In fact, this would also explain why a trending significant correlation was found between the Plausibility difference wave in this time window and RTs, but not for Order. Due to these observations, we note that the plausibility effect seems robust, but the sustained negativity in implausible TD trials after 500 ms should be taken with caution.

Differences in the later 650-800 ms window showed that there was also a main effect of Order, for both plausible and implausible trials. This effect was significant in left posterior electrodes and largest in right frontal areas, with a significant negativity found in distractor-last trials (TD) relative to target-last (DT). This Order effect is in line with the assumption that the Target-Distractor trials pose an increased demand for controlled selection, given that participants must discard the most recently (and therefore strongly) activated candidate. However, it is only partly congruent with the LPC reported in previous studies. While it matches the inverted negativity for conflict conditions in frontal electrodes, the elicited wave is also more negative for TD trials; that is, the polarity of the effect does not differ in left posterior electrodes.

Given these predictions and the topography of this effect, we interpret it in connection with the literature on memory and retrieval. The functionality and scalp distribution of this negativity seem to be congruent with the left parietal effect and right-frontal effects (RFE) of memory retrieval studies conducted using the old/new paradigms. The left parietal old/new effect elicits a more positive wave for correctly retrieved items, and
a more negative wave in correct rejections (Wilding, 1999, 2000); however, other studies have reported mixed results, with similar amplitudes for inaccurate recollections (e.g., Senkfor & Van Petten, 1998; Van Petten & Senkfor, 2000). Although the functionality of these components is still debated, it has been suggested that it is associated with memory retrieval. Relevant to the hypotheses tested here, some studies have found a correlation between the amplitude of the effect (referred to as LPC\textsuperscript{11} in some studies, e.g., Finnigan, Humphreys, Dennis & Geffen, 2002) and response accuracy, in line with our own data. Importantly, it has been proposed that this component reflects decisional factors during memory retrieval, including accuracy and possibly confidence in the response. The similarities with the current paradigm are quite apparent, given that distractor-last trials require the evaluation and monitoring of the accessed representation (distractor – noun) and its rejection based on the phrases presented earlier during the study.

However, as noted, the effect was present most prominently in right frontal electrodes in addition to left posterior electrodes. The right-frontal old/new effect (RFE) has been associated with “post-retrieval” monitoring processes (e.g., Allan, Wilding & Rugg, 1998; Hayama, Johnson & Rugg, 2008; Leynes & Kakadia, 2013). The RFE is associated with monitoring of retrieved episodic information while a judgement is made (i.e., post-retrieval monitoring, Wilding, 1999; Hayama, Johnson & Rugg, 2008; Leynes & Kakadia, 2013). Similarly, here distractor-last trials required the evaluation and monitoring of the accessed representation (distractor – noun), and their rejection based on the phrases presented earlier during the study.

\textsuperscript{11} A clarification about the labels of components seems necessary here. As Finningan and colleagues point out, the terminology is inconsistent across studies within the ERP literature on retrieval, with the same component sometimes being referred to as LPC, P3, P300, or P600. More confusingly, the same label (LPC) refers to a positivity in known items in these studies for which retrieval is successful, and is believed to index a matching representation, while in the LPC in the Stroop literature, the positivity is associated with a cost in selection for incongruent (conflict) trials. In other words, while the literature on retrieval emphasizes the polarity of trials with successful recollection (positivity), the conflict literature describes the polarity of conflict trials (negativity).
Crucially, the results of the analysis, showing that the difference waves between 650 and 800 ms correlated with accuracy, support the interpretation of this negativity as reflecting monitoring processes during selection. Given that the peak of the effect was present in right frontal electrodes, and in order to follow previously employed terminology, we will refer here to this right frontal late negativity as the right frontal effect (RFE), while emphasizing the association of a negativity with a cost in retrieval during selection.

All in all, the effects identified across the three time windows allow us to differentiate the aggregated effects reflected in the behavioral data, providing a finer-grained analysis of the qualitative costs incurred during lexical selection. To summarize, consistent with RT results, the cost of processing the noun in implausible TD trials, as well as the cost associated with distractor-last trials, accounts for the significantly slower RTs relative to implausible DT. Plausible trials elicited a conflict-detection component (the Ninc) within the 500-600 window. Finally, plausible TD trials showed an Order effect in the negativity from 650-800 ms, maximal over right frontal electrodes, as well as the longest RTs in the behavioral results.

**ERP analysis of Spanish-English bilinguals**

The behavioral results reported above showed that averaged RTs started shortly after 950 ms in bilinguals; that is, there was a delay of about 150 ms relative to native speakers. Therefore, the mass univariate analysis was adjusted accordingly to examine the time window from 500 to 950 ms. Because extending the time window of the exploratory analysis involves an increased loss of statistical power due to more comparisons being made, the analysis was restricted to the electrodes of areas of interest; these were the left centroparietal and right frontal areas of the native English speakers, where the effects were expected.

For the non-native speakers, running t-tests revealed differences modulated by Plausibility over left posterior electrodes (CP1, CP5, P3, P7). A sustained effect was found from 600 to 800 ms, with more irregular intervals of significance across the same electrodes between 800 and 900 ms (significant intervals are illustrated for CP1 by gray bars in Figure 3-
6; the supplementary materials show the differences from running t-tests for each significant electrode). The fact that the significant effect appeared later is congruent with the delay observed in RTs. Unlike in the native speaker group, no other differences were revealed. Nonetheless, based on the late right frontal negativity (RFN) found in the native group during the final 150 ms preceding the beginning of RTs, and for comparability between native and non-native speakers, it was deemed worth investigating this final time window. The ERP analysis therefore considered three time-windows: 300-500 (N400), 600-800 ms and 800-950 ms.

**N400 component (300-500 ms)**

The midline ANOVA revealed main effects of Order ($F(1, 20) = 6.56, p < .05$) and Frontality ($F(2, 42) = 7.72, p < .01$), and an interaction of Plausibility x Order ($F(1, 20) = 5.0, p < .05$). Similarly, the lateralized ANOVA also showed main effects of Order ($F(1, 20) = 7.78, p < .05$) and Frontality ($F(2, 42) = 11.71, p < .0001$), as well as interactions of Plausibility and Order ($F(1, 20) = 6.32, p < .05$) and Order x Hemisphere x Frontality ($F(1.05, 20.91) = 4.1, p < .05$). Pairwise comparisons showed that, crucially, implausible TD trials ($EAT – shoot – breakfast$) were different from all other conditions over left and right posterior electrodes (all $p < .0001$), with no other significant differences. In frontal electrodes, all comparisons showed significant differences except for plausible DT vs plausible TD.

**600-800 ms window**

The results of the midline ANOVA revealed main effects of Plausibility ($F(1, 20) = 5.12, p < .05$) and Frontality ($F(2, 40) = 4.6, p < .05$). In the lateralized analysis, there were main effects of Plausibility ($F(1, 20) = 5.09, p < .05$) and Frontality ($F(2, 40) = 5.8, p < .01$), and interactions between Order x Hemisphere ($F(1, 20) = 4.57, p < .05$) and Hemisphere x Frontality ($F(2, 40) = 6.9, p < .01$). Three-way interactions emerged between Plausibility x
Hemisphere x Frontality ($F(2, 40) = 3.41, p < .05$), and Order x Hemisphere x Frontality ($F(2, 40) = 71.91, p < .0001$). The three-way interactions suggest that the main and effects of Order and Plausibility differed across scalp regions. Pairwise comparisons confirmed that, in left and right posterior electrodes, differences were significant only between plausible and implausible trials, with no effect of Order. This was also the case in left frontal electrodes, although comparisons for plausible TD were marginally significant with plausible DT ($t(83) = 1.99, p = .07$) and not significant with implausible DT ($t(83) = -1.57, p = .11$). Over right frontal electrodes, implausible DT (shoot – EAT – breakfast) was different from all other conditions ($p < .01$), with no other differences. Difference waves in Figures 3-7a-b show the scalp distribution of the effects, with a widely distributed effect of Plausibility (3-7a).

800-950 ms window

The midline ANOVA only showed a significant effect of Frontality ($F(2, 40) = 11.36, p < .0001$). In the lateralized analysis, significant main effects were found for Frontality ($F(2, 40) = 9.0, p < .001$) and Hemisphere ($F(2, 40) = 5.77, p < .05$). There were significant interactions of Order x Hemisphere ($F(1, 20) = 9.63, p < .01$), Hemisphere x Frontality ($F(2, 40) = 6.88, p < .01$), Order x Hemisphere x Frontality ($F(2, 40) = 54.31, p < .0001$) and Plausibility x Order x Hemisphere x Frontality ($F(2, 40) = 4.69, p < .05$). Pairwise comparisons showed the same effects as in the previous time window for left posterior electrodes, with significant differences only between plausible and implausible conditions, and no effect of Order. Over right frontal electrodes, once again implausible DT (shoot – EAT – breakfast) was less negative than all other conditions ($p < 0.001$).
Figure 3-6: ERP waveforms for non-native speakers at (a) FC6 (right frontocentral) and (b) CP1 (left centroparietal). The legend shows examples for each condition, with the target verb capitalized. Grey bars underneath show the time points of significant effects in running t-tests within the pre-defined 500-950 ms time window. Boxes show the three time windows analyzed. The arrow in (b) indicates the N_{inc} effect. Negativity is plotted up.
Correlation between neurophysiological and behavioral data

Based on the results of the experiment with native speakers, we performed one-tailed correlations between individual ERP amplitude differences waves and the behavioral results. Differences between plausible and implausible DT trials within the 600-800 ms and 800-950 ms time windows were not correlated with accuracy nor RTs. Correlations between the individual familiarity ratings for multiword units and amplitude difference waves were also non-significant.

Discussion of ERP results in non-native speakers

We used the same materials and paradigm as with native speakers to investigate the ERP components associated with conflict detection and resolution in late Spanish-English bilinguals. The results revealed that the plausibility manipulation produced similar effects in
non-native speakers of English, but there were also some notable differences. The behavioral results showed similar patterns to those found in native speakers. However, in both the analysis of accuracy and in RTs, only a main effect of Plausibility emerged; there was no significant main effect of Order. In the analysis of accuracy, the interaction of Order x Plausibility was significant, showing that non-native speakers made more errors when distractors were plausible and were presented last (EAT – skip - breakfast). Additionally, unlike in native speakers, the measure of English Verbal Fluency (which accounted for individual language ability in English) was a significant predictor of overall performance in the task.

In the ERP analysis, the N400 component showed the expected cost in processing the noun in implausible TD trials (EAT – shoot – breakfast). Interestingly, a significant difference emerged in the right posterior region between plausible TD and implausible TD, but not between plausible TD and DT. This suggests that processing the noun did not seem to significantly differ between trials in which it was immediately preceded by a plausible distractor and trials in which it was preceded by the target.

The analysis of the 600-800 ms window showed a widely distributed effect of Plausibility in the comparison between plausible and implausible target-last trials. As shown in Figures 3-7a-b, the effect was widely distributed; however, there was no effect of Order, as shown in Figures 3-7c-d. Over right frontal electrodes, the baseline condition significantly differed from all others, showing a generalized cost of both Order and Plausibility. The absence of an Order effect in the comparison between plausible TD-DT, with both showing a negativity relative to the baseline, suggests that rather than no cost in distractor-last trials, there was simply no advantage in target-last trials. That is, the results suggest that lexical selection in trials with plausible distractors produced a generalized cost for non-native speakers. The absence of a facilitation in DT trials appears to be caused by an increased need for monitoring during retrieval in bilinguals.

Additionally, a more generalized cost was widespread across scalp locations (Figure 3-7a-b). More importantly, unlike in native speakers, no two components could be separated,
although differences in the effect over right frontal electrodes became more highly significant between 800 and 950 ms. The impossibility to temporally separate different components may be due to a greater degree of variability among non-native speakers, i.e., in their proficiency or other individual differences regarding the L2. Even though we restricted our sample to late non-simultaneous Spanish-English bilinguals immersed in their L2, processing in the non-native language is known to be more variable and other aspects of their experience and learning might have introduced additional variables. However, an alternative explanation is possible. We suggest that, in non-native speakers, processing may be more costly in terms of retrieval and monitoring, demanding engagement of more cognitive resources. Therefore, there may be increased overlap in the timing of conflict detection and monitoring in bilinguals. We elaborate on this idea in the general Discussion.

In summary, the ERPs of bilingual participants showed a main effect of Plausibility, in line with the behavioral data. The convergence of behavioral and ERP data suggests qualitative differences in the bilingual group relative to native speakers, in that Plausibility but not Order seemed to pose a cost in selection. We turn to this question in connection with similar prior evidence in bilinguals in the last section.

Discussion

The present study examined the cognitive mechanisms involved in conflict resolution in native and non-native speakers during lexical selection in a context where selection takes place in units larger than a single word (e.g., eat breakfast). To do so, we employed a paradigm with sequential presentation of two verbs and a noun (e.g., \textit{EAT} – \textit{skip} – \textit{breakfast}). The materials were manipulated so that distractors would either be plausible in English (as in the previous example) or implausible (\textit{EAT} – \textit{shoot} – \textit{breakfast}). Importantly, in the case of non-native speakers (i.e., late Spanish-English bilinguals) the plausible distractor + noun combinations formed multiword units that were incongruent with bilinguals’ native language (\textit{skip breakfast} is “saltarse el desayuno”, ‘to jump the breakfast’). The order of presentation of
the distractor verb and the target verb was counterbalanced (EAT – skip or skip – EAT). This 2 x 2 design allowed us to examine (a) how the verb that was presented most recently affected processing of the noun (i.e., priming), during the N400 window; (b) the conflict-related ERP components that are elicited in high conflict contexts (both verbs are plausible) relative to low conflict contexts (implausible distractor); (c) the ERP correlates of conflict-resolution and monitoring prior to response selection.

The results of the experiment showed that processing of the noun was facilitated by priming from the verb in the native speaker data when the target immediately preceded the noun (eat breakfast), but also when a matched plausible distractor primed the noun (skip breakfast). In both datasets, ERPs time-locked to the presentation of the noun showed an attenuated N400. The same pattern was found in the bilingual speaker dataset. Interestingly, neither the analysis of the native nor non-native speakers’ data showed significant differences between trials in which the noun was primed by the target verb (e.g., EAT) or by the plausible distractor (skip). This is remarkable because, for bilinguals, target verbs + nouns formed multiword units congruent across the L1-L2, while plausible distractors + nouns formed L1-L2 incongruent multiword units.

Taking direction from the psycholinguistic literature on conflict detection, we expected that the Ninc, a conflict-related component described in the ERP studies using the Stroop paradigm, should emerge in the time period after processing of the noun and before a response was made. Results for native English speakers showed a negativity (500-600 ms), maximal in left centroparietal electrodes. This effect was congruent with the functionality and scalp distribution of previous studies that identified the Ninc using the Stroop paradigm. The negativity was modulated by Plausibility, showing that interference from a valid candidate elicited the Ninc component; it was also present when conflict arose from the need to discard an implausible target that was presented immediately before the noun. The fact that the Ninc would be elicited for different types of conflict is in line with the evidence that conflict-detection is a domain-general mechanism, elicited by linguistic and non-linguistic conflict (i.e., visual conflict in direction, as in the Flanker task; semantic conflict mediated by
modality, as in Stroop; Larson, Clayson & Clawson, 2013). Additionally, a marginally significant correlation between ERP difference waves and RTs suggested an association between conflict detection and the delay of responses.

In addition, monolinguals showed a later component before responses were made. The functionality and scalp distribution of this negativity were congruent with the right-frontal old/new effect (RFE), believed to reflect general “post-retrieval” monitoring processes (e.g., Allan, Wilding & Rugg, 1998; Hayama, Johnson & Rugg, 2008; Leynes & Kakadia, 2013; Rugg & Curran, 2007; Rugg, Herron & Morcom, 2002). The effect is associated with monitoring of retrieved episodic information while a judgement is made (i.e., post-retrieval monitoring, Rugg, Herron & Morcom, 2002). The results of the correlations showed that the difference waves between 650 and 800 ms correlated with accuracy; this supports both the interpretation of this negativity as the RFE as well as the proposal that this component reflects post-retrieval monitoring that supports accurate selection. The component was observed to be modulated in native speakers by Order affecting the selection of the target when verbs competed for selection. Importantly, the analysis showed that the amplitude of the RFE was significantly correlated with the proportion of accurate responses. This provides compelling evidence that this potential is associated with monitoring during goal-oriented selection of the target.

Results for bilingual speakers revealed important similarities but also some different patterns. First, similarly to the analysis of the native speaker data, the bilingual data showed that plausible verbs primed the noun within the 300-500 ms time window, and this was the case for both target verbs (eat breakfast) or plausible distractors (skip breakfast), despite the cross-language differences for non-native speakers. Second, preceding the selection of the target verb, ERPs also showed significant differences modulated by Plausibility. Similar to the native speaker data, the effect was maximal over left centroparietal and right frontal electrodes. However, there were also some notable differences regarding the timing of selection-related components. While in native speakers an N100 was followed by a right frontal effect (RFE), indicating a cost in post-lexical retrieval, in non-native speakers, the timing of
both effects was indistinguishable. Maxima were found over left centroparietal and right frontal electrodes in both groups, but in non-native speakers the effect manifested in the form of a more sustained and widespread negativity. This conflict-related negativity was more prolonged in time (600-800 ms), appeared to amalgamate both the \( N_{inc} \) and RFE, and became more prominent in right frontal electrodes shortly before responses were made (800-950 ms).

We suggest that the findings of the non-native speakers are in line with our current understanding of cross-language activation in bilinguals. On the one hand, it is not unexpected that non-native speakers should experience increased demands for monitoring and selection. The need for sustained or prolonged monitoring would be justified by the fact that bilinguals activate more information than monolinguals, including activation from multiword units in the L1 (Siyanova-Chanturia, Conklin & Schmitt, 2011; Carrol & Conklin, 2017). While we refrain from making any specific claims regarding the brain areas engaged in the two populations tested, there is good evidence from neuroimaging studies that using the weaker language results in greater and more widespread brain activation in bilinguals (for a recent review, see Abutalebi & Green, 2016). For instance, a “prefrontal effect” is associated with usage of the less dominant language (Abutalebi & Green, 2007). These cortical regions are not involved in conflict detection but in the process of response selection and suppression (Green & Abutalebi, 2013). Additionally, increased activation in the left inferior parietal lobule is associated with biasing lexical selection towards the language in use (Price, Green & von Studnitz, 1999). In light of this evidence, it is not surprising that bilinguals would show a more generalized cost, manifested in the widespread effect revealed in our data. Also, because the effect was more extended in time, this might suggest a need to maintain and re-evaluate competitors in memory. Thus, the overlapping of monitoring and selection components may reflect a processing cost due to less entrenched representations of multiword units in the L2 and uncertainty in the correct lexical choice. This would be in line with some accounts suggesting that the RFE may indicate a greater number of computations required in response selection (Hayama, Johnson & Rugg, 2008; Dobbins & Han, 2006). The current data does not
allow us to determine the loci of the costs, and these interpretations will need to be tested in future research.

Further, beyond the timing of the processes, the significance of the effects also showed both similarities and differences in behavioral and ERP data. Behaviorally, there was a main effect of Plausibility in the RTs of both native and non-native speakers, as well as a main effect of Order. However, the patterns found in the accuracy data revealed that the effect of Order was robust only in the native speakers, who consistently showed a cost in distractor-last trials. In contrast, non-native speakers did not show a reliable effect of Order. In fact, when implausible distractors were presented last, bilinguals showed an advantage in selection. These patterns in the accuracy data are indicative of differences in the processing or selection strategies across groups, with Order having a more consistent effect in native speakers, while bilingual speakers are consistently affected by Plausibility. While the effect of Group was tested directly only in the behavioral analysis, we note that the by-group ERP results were consistent with these patterns. In native speakers, the conflict-detection component over left centro-parietal electrodes showed a negative peak for trials with plausible distractors as well as for target-last trials; in non-native speakers, left centroparietal electrodes showed a significant effect of Plausibility but not Order, in line with the suggestion that conflict detection was not modulated by Order in bilinguals.

Why would plausibility, rather than order, have a greater effect in non-native speakers than in native speakers? The answer may lie in known differences between processing in the L1 and the L2 in general and, more specifically, in processing of multiword units by adults. Current theories of sentence processing, such as the Shallow Structure Hypothesis (SSH, Clahsen & Felser, 2006a, 2006b, 2018), posit that non-native speakers often achieve native-like performance in online processing tasks, but do so by relying on non-grammatical (e.g., lexical) cues to a greater extent than on syntactic and morphosyntactic cues. To illustrate, evidence from online processing of syntactic subject/object ambiguities has shown that non-native speakers first rely on lexical and plausibility information, whereas reliance on plausibility information is typically delayed or absent in native speakers (Roberts
& Felser, 2008). Therefore, the evidence suggests that when it comes to lexico-semantic interpretations, non-native speakers have greater sensitivity to plausibility.

An alternative, yet compatible, interpretation is based on the fact that monolinguals have entrenched knowledge of multiword units, and are sensitive to variations in form, i.e., expect to encounter them in a particular form. As mentioned in the introduction, previous research has shown faster processing in native speakers when a multiword unit is presented in its exact familiar form and order, e.g., “black and white” relative to “white and black.” Based on these findings, the order of the verbs before the noun may have played an important role in how monolinguals approached the task, potentially focusing on recognizing the expected verb-noun sequence. In addition to this, evidence has shown that L2 speakers are successful at learning individual words but they have difficulty in acquiring multiword units (McCauley & Christiansen, 2017; Nguyen & Webb, 2017). Indeed, it has been proposed that it is precisely non-native speakers’ greater focus on individual words and on decomposing multiword units that poses a constraint in acquiring L2 multiword units (Arnon & Christiansen, 2017). The discussed focus of non-native speakers on semantics, as well as the lesser reliance on structural cues, might have likely affected their approach to the task, with increased Plausibility-based costs, while relegating Order-based factors to a secondary stage. These factors may have together or independently contributed to the greater Order-based cost found in native speakers than in bilinguals.

Conclusions

In the experiments presented, we investigated the brain potentials associated with conflict detection and resolution during multiword lexical selection. The novel paradigm employed allowed us to go beyond selection at the single word level to examine how, in forming everyday phrases, speakers detect and resolve conflict between the word choices available to them. When native and non-native English speakers were asked to select a familiarized target Verb-Noun sequence (eat breakfast), we observed that the context of selection modulated
brain potentials. Specifically, contexts that resulted in conflict during selection due to more than one valid candidate induced negative-going potentials, as well as contexts in which the most recently processed verb had to be rejected (e.g., eat – skip – breakfast; target: eat).

However, native English speakers seemed to experience a greater cost from Order of presentation (i.e., when the consecutively presented Verb-Noun sequence was not identified as the target). On the other hand, ERP components in late Spanish-English bilinguals were most significantly modulated by Plausibility. In other words, native speakers experienced a cost from not encountering the expected sequences of multiword units, while selection costs in non-native speakers were most affected by the plausibility of each individual combination.

The bilingual finding is in agreement with current theories of processing that propose that, while native speakers rely on multiword units, non-native speakers processing is more focused on individual words (Arnon & Christiansen, 2017; McCauley & Christiansen, 2017; Nguyen & Webb, 2017).
Chapter 4

Connecting practice conditions, neurophysiology and learning: training native language inhibition improves L2 lexical retrieval

Introduction

The main goal of this experiment is to investigate the neurophysiological correlates of “retrieval practice” in “desirably difficult” conditions, akin to those that were found to improve learning in the first experiment (Chapter 2). By using EEG, our goal is to identify the specific mechanisms associated with practice conditions that improve learning outcomes.

The results from Experiment 1 provided evidence in support of the hypothesis that native language regulation may play an important role in learning and in the ability to retrieve L1-L2 incongruent multiword units. More specifically, it showed that inhibition of the L1 counterparts of L1-L2 incongruent collocations was associated with higher rates of recall in a translation test. Informed by the notion that L1 inhibition might play an important role in L2 learning, in Experiment 2 we further investigated the mechanisms associated with lexical selection in cases in which one competitor must be suppressed. Taking direction from the voluminous work in the literature on conflict detection and resolution, including EEG studies (e.g., Coderre, Conklin & van Heuven, 2011; Glaser & Glaser, 1982; Liotti, Woldorff, Perez & Mayberg, 2000; West & Alain, 1999; West, 2003), we hypothesized that conflict-related ERP components might provide a neurophysiological index associated with the ability to regulate the L1 during L2 learning. The results from the lexical selection task in Experiment 2 showed that the N_{inc} component, considered a conflict-detection component, could be observed in native speakers but was not detected in non-native speakers. Further, both groups presented a right frontal effect (RFE), manifested as a negative-going wave in high-conflict trials which required greater monitoring during response selection. Although this component was correlated with accuracy in the responses of native speakers, the RFE was not found to correlate with accuracy in non-native speakers. The lack of a significant association between
ERPs and behavioral responses in the non-native group could be attributed to greater variability in non-native speakers and, perhaps, to a potential overlap of monitoring and selection components. Taken together, the results of these studies would suggest that L2 lexical selection in non-native speakers is associated with brain potentials that differ from those found in the L1. Thus far, inhibition of L1 competitors stands as the most likely correlate in successful learning of L2 multiword units, but no evidence in support of this view has been provided beyond the behavioral level. The main goal of this third experiment is to use a paradigm analogous to the one used in Experiment 1, to test the idea that L1 inhibition can unequivocally predict L2 attainment. To this end, we investigated whether ERPs elicited during different practice conditions can predict performance during subsequent testing.

Additionally, this experiment will allow to answer three additional questions that stem from the findings reported in Chapter 2 and Chapter 3: (a) whether significant differences are found when practice conditions engage other modalities of response selection (i.e., manual responses instead of oral production); (b) how using tests that pose different demands on retrieval affect rates of accurate recall; and (c) whether the benefits in interference suppression derived from practice in desirably difficult conditions are generalizable to contexts that require regulation of lexical competitors within- or across-languages. In what follows we elaborate on these goals.

**The effect of practice modality during lexical selection on learning**

This experiment is a conceptual replication of Experiment 1, inasmuch as it employs the same experimental manipulation in the design of the materials, and a similar experimental procedure. However, because of some methodological considerations in ERP studies, the Practice procedure of Experiment 1 (Chapter 2) has been adapted here by following the same response selection method described for Experiment 2 (Chapter 3). That is, we implement a substantial modification in the mode of response for the Practice on lexical selection relative to Experiment 1, in that instead of providing oral responses, learners selected the target by
pressing a button. This was done for two different reasons. First, pressing buttons rather than spoken language responses helps avoid motor artifacts derived from speech. While it has been shown that ERP data can be analyzed even in the presence of such artifacts (Ouyang, Sommer, Zhou, Aristei, Pinkpank & Abdel Rahman, 2016), relying on manual responses circumvents the potential methodological challenges in testing this experimental manipulation using EEG for the first time. Secondly, and perhaps more importantly, conducting the experiment with manual rather than vocal responses will allow us to investigate the role of mode of practice in replicating the results of Experiment 1. Research has suggested that engaging the production system may play an important role in learning, through what is known as the “production effect” (Forrin & MacLeod, 2017; Ozubko & MacLeod, 2010; Zormpa, Brehm, Hoedemaker & Meyer, 2018). On the one hand, overt responses feed into the auditory loop and have greater impact when produced by oneself rather than by another person (Forrin & MacLeod, 2017). On the other hand, overt speech engages the production system, encompassing several levels of retrieval, including gesture and motor planning. The available evidence suggests that production may play an important role in learning by enforcing retrieval of form-meaning associations at multiple levels of representation, with enduring effects even after several days (Ozubko, Hourihan & MacLeod, 2012). Given this, the inclusion of vocal responses in Experiment 1 may have provided the optimal practice conditions to test the effectiveness of our manipulation. However, in language learning exercises and testing, students are regularly asked to select a correct choice by making a response through the click of a computer mouse or by circling the correct option on a piece of paper. While manual responses are certainly an ecologically valid form of response selection, the evidence on the production effect suggests that such non-verbal responses may be less effective. The question of whether non-vocal responses are equally effective as a mode of retrieval may be particularly relevant at this time, given the fact that several factors (e.g., the shift towards hybrid instruction) have contributed to diminishing the number of opportunities for overt vocal responses in practice. Therefore, the question of whether modality of lexical selection affects learning is one of both theoretical and practical import.
The effect of test format on retrieval accuracy

A potential limitation of Experiment 1 is that recall of learned multiword units was measured using a translation test only. Despite the advantages associated with translation tests (they directly assess the ability to retrieve L1-to-L2 mappings, and are as restricted as possible with regards to what target response is expected), translation is not without its limitations in that it requires identifying an explicit cross-language equivalence and may not be representative of lexical retrieval in some real-world situations. Based on this, the current experiment also included a cloze test, which presented participants with sentences containing the learned collocations in a L2-only context, with a gap to be filled using only the appropriate verb.

Exploring the generalization of benefits derived from desirable difficulties

An additional goal of this experiment is to explore whether, and how, benefits derived from practice in desirably difficult conditions are generalizable. In other words, are the benefits of training in interference suppression limited to the items studied? Or does practice in this skill generalize to other contexts requiring conflict resolution? The present experiment includes two pathways to investigate this question. First, a second list of L1-L2 incongruent collocations was included, which was practiced in identical ways by all learners, allowing us to test whether other aspects of training have any consequences that carry over to the L1-L2 incongruent collocations not subject to any experimental manipulation. This aspect of the experiment would provide an opportunity to test generalization of the ability to regulate cross-language interference during learning. Secondly, the cloze test included two degrees of context constraint: high- and low-constraint carrier sentences. While high-constraint sentences restrict the number of competing alternative responses, low-constraint sentences have the opposite effect, requiring learners to more effectively block off non-target competitors in order to provide the correct response. Therefore, this manipulation of the cloze
test allows us to investigate generalization in the ability to regulate *within-language* interference.

The goals of the study are formulated in the research questions below, which are followed by our predictions.

**Research Questions**

1. What ERP components are elicited by Practice in conditions of low interference (unrelated distractors) or high interference (when L1-related distractors are presented)?
2. Is the neurophysiological activity elicited during Practice predictive of recall rates in post-tests?
3. Does Practice that only requires manual selection during retrieval significantly improve learning?
4. How do the effects of Practice differ when measured by Translation and Cloze post-tests?
5. Do any potential benefits derived from different Practice conditions generalize to recall of other collocations (trained in the same conditions across groups)?

**Predictions**

1. Based on the findings in previous studies that used a highly comparable paradigm, we hypothesized that retrieval Practice requiring rejection of L1-related distractor verbs would result in inhibition of those verbs (Experiment 1, Chapter 2). We predicted that the effects of inhibition would be manifested in a greater N400 amplitude for the nouns of congruent (rather than incongruent) collocations during the Practice. Likewise, based on previous ERP evidence from a similar multiword selection
paradigm with non-native speakers (Experiment 2, Chapter 3), we predicted that a right frontal effect (RFE) might be observable between 600 and 800 ms.

2. Because previous behavioral findings have demonstrated a direct correlation between inhibition of L1-related distractors and higher recall rates in a post-test, we expected that ERPs indexing inhibition might show the same association. This association was expected to be particularly strong in the Translation test used in the current study, which requires more direct suppression of the L1 equivalents presented and is comparable to previous data (Experiment 1, Chapter 2).

3. It remained an empirical question whether the non-vocal Practice employed here, in which participants select verbs through button responses, would be as effective as training requiring overt vocal responses. However, given previous evidence that producing overt responses may engage different levels of representation and improve retrieval ability, we hypothesized that the non-vocal practice design employed in the present study might be less effective.

4. The Translation test, which also presented the L1 equivalents, was expected to rely on language regulation to a greater extent than the Cloze test, potentially eliciting a stronger effect of Practice group.

5. If any retrieval mechanisms were enhanced by the Practice conditions, we hypothesized that any benefits in the incongruent collocations subject to manipulation might generalize to incongruent collocations trained in the same conditions across groups.

Methodology

Participants

A group of 61 undergraduate learners of English was recruited at a large US university. Participants were native speakers of English completing a third semester
university-level elementary Spanish course (roughly equivalent to the levels A2-B1 of the Common European Framework of Reference for Languages, Council of Europe, 2011). A sample size of 21 participants per condition was pre-determined based on a power analysis of data from Experiment 2 (Chapter 3) for an alpha threshold < .05 (power: 0.90) and was also based on a similar sample size in Experiment 1 (Chapter 2). All participants gave informed consent and were paid 10 US dollars per hour of participation. To confirm their eligibility, participants completed a collocations baseline test designed to ensure that participants were knowledgeable of the individual words that composed the collocations, but were not familiar with the collocations per se. Eleven participants were excluded after the first session due to prior knowledge of the collocations. Data from three additional participants were excluded due to experimental error. The remaining participants (N = 47) were randomly assigned to one of two learning conditions: a group in which unrelated distractors were presented during retrieval practice (henceforth, the ‘Unrelated’ group, N = 24; 63% female), and a group that saw L1-related distractors during recall of a list of incongruent collocations (‘L1-Interference’ group, N = 23; 52% female).

**Individual differences measures**

Participants completed a number of measures of language proficiency in the L1 and L2, as well as additional tasks to ascertain individual differences in cognitive control. Here we briefly enumerate the measures and report the results below (see Table 4-1).

To assess linguistic profile and background in the L1 and L2, participants completed the LEAP-Q (Marian, Blumenfeld & Kaushanskaya, 2007). Our selection of other cognitive and proficiency tasks was informed by measures found to be significant predictors in the first experiment (Chapter 2), which include the Flanker task (Eriksen & Eriksen, 1974) as a measure of cognitive control and the Nonword Repetition task (Baddeley, Papagno & Vallar, 1988) as an index of short-term phonological memory (PSTM). Lexical knowledge and fluency was measured through the Category Verbal Fluency. Two versions of the task were
administered, one in English first and then in Spanish, each containing two categories. Participants also completed a multiple-choice test to assess knowledge of the individual words employed in the experiment, which they were expected to know. In addition, they completed a baseline collocations pre-test (reported in Section “Test administration, scoring and analysis”) to ensure that participants were not familiar with the collocations included in the experiment. All participants were right-handed, as confirmed by the Edinburgh handedness test (Oldfield, 1971). Groups were matched across all measures.

Table 4-1: Summary of cognitive and proficiency measures.

<table>
<thead>
<tr>
<th></th>
<th>Unrelated condition (N=22)</th>
<th>L1-interference condition (N=23)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valid N</td>
<td>M  SD</td>
</tr>
<tr>
<td>Age (in years)</td>
<td>22</td>
<td>19  1.18</td>
</tr>
<tr>
<td>L2 OoA (in years)</td>
<td>22</td>
<td>12.55  3.1</td>
</tr>
<tr>
<td>L2 self-rated proficiency (/10)</td>
<td>22</td>
<td>4.18  2.07</td>
</tr>
<tr>
<td>Weekly exposure to L2 (%)</td>
<td>22</td>
<td>5.21  4.62</td>
</tr>
<tr>
<td>English (L1) Verbal Fluency</td>
<td>22</td>
<td>30.08  6.46</td>
</tr>
<tr>
<td>Spanish (L2) Verbal Fluency</td>
<td>22</td>
<td>9.5  4.06</td>
</tr>
<tr>
<td>Baseline Vocabulary Test (/100)</td>
<td>22</td>
<td>82.3  5.33</td>
</tr>
<tr>
<td>PSTM: Nonword repetition (/100)</td>
<td>21</td>
<td>60.63  9.34</td>
</tr>
<tr>
<td>Flanker effect (ms)</td>
<td>22</td>
<td>49.44  27.27</td>
</tr>
</tbody>
</table>

**Materials**

Three types of materials were created, following a design and procedure similar to Experiment 1 (Chapter 2): materials for the Familiarization phase, for the Practice phase and for the Testing phase. The materials for Familiarization phase and the Testing phase were
identical for all participants, while the materials used for Practice differed across the two groups. We first describe the list of collocations used to create the experimental materials, and subsequently describe the materials and procedure for the Familiarization, Practice and Testing phases.

**List of collocations**

Forty-two collocations were extracted using the web-based version of Corpus del Español (CdE, Davies, 2016), with over 2 billion words, and data for equivalent English collocations were extracted from the Contemporary Corpus of American English (Davies, 2008), with over 560 million words. The stimuli contained three types of collocations (see Table 4-2 for a sample of the materials). One first list contained fourteen L1-L2 incongruent collocations in which the noun, but not the verb, was equivalent across Spanish and English. For example, the Spanish collocation “gastar bromas” is conventionally expressed in English as “play jokes”, although it would literally translate as ‘spend jokes’. We will refer to this as the ‘Experimental Incongruent’ list. A second list contained an equal number of ‘L1-L2 Congruent’ collocations derived from the first list. We first identified the idiomatic equivalents in English for the first list (“play jokes” in the example above) and used the verb in the literal translation to search for an L1-L2 congruent collocation (e.g., for “play” – “jugar”, we selected “jugar partidos”, which means ‘play matches’). This way, the collocations in the first two lists were associated by virtue of this manipulation. Finally, a third type of “independent incongruent collocations” was created (e.g., “marcar un número” ‘dial a number’). This list was included to test for the effect that different practice conditions in each group might have on the ability to learn and retrieve an additional independent list of L1-L2 incongruent collocations. In order to distinguish the two lists of incongruent collocations, we will refer to this as the ‘Independent Incongruent’ list, in contrast to the first ‘Experimental Incongruent’ list described above.
Words were not repeated across any of the items in the lists, which were matched on log frequency, collocational strength, noun syllable length and orthographic length. To determine the collocational status of verb-noun phrases, t-scores were used as the statistical association measure (all t-scores > 4.0). Half of the nouns in each list were cognates. The complete list of stimuli is available in Appendix E.

**Experimental Procedure**

The study was divided in three sessions. The first two lasted approximately two hours and were completed in the same week; the third session, in which the delayed post-tests were administered, was completed about a week later (mean: 5.61 days, SD: 0.21). In the first session, participants completed the battery of cognitive and proficiency tests. The second session consisted of the Familiarization phase, Practice task and Tests. Here we succinctly describe the materials and procedure for each task.

**Familiarization materials**

Based on the lists of materials, stimuli were created for visual and auditory presentation of the target Spanish collocations and their idiomatic English equivalents. The Spanish stimuli were recorded by the first author and their English counterparts were recorded by a native English speaker.

**Familiarization procedure**

The Familiarization phase was presented using E-Prime 3.0 (Psychology Software Tools, Pittsburgh, PA) in the following manner. First, a screen showing the Spanish collocation was presented, followed by the English collocation; the corresponding audio recording was played simultaneously for each screen. Next, participants were instructed to
repeat twice the target Spanish collocation: first by saying it out loud into a microphone and, on the following screen, by typing the Spanish verb-noun collocation. Feedback was immediately provided, such that if the response was typed perfectly, the same collocation was displayed in blue font; if the response typed did not exactly match the target, the correct response was shown in red font (Figure 4-1).

To ensure sufficient input, the same forty-two collocations were shown again by administering a second round. This time, in order to ensure that both the Spanish phrases and their meaning were learned, participants were also prompted to remember and type the English meaning of the Spanish collocations at regular intervals, after every seven trials. The entire familiarization procedure took approximately 25 minutes.

Figure 4-1: Sequencing in the Study procedure.
Practice materials

Three sets of Practice materials were created. One set contained all the items in the ‘Independent Incongruent’ collocations list. As indicated above, this list was practiced in identical conditions in both groups of participants, and it was included to test for whether different practice conditions across groups might affect the ability to learn and retrieve a third additional independent list of L1-L2 incongruent collocations. In this set, verbs from the same list were used as distractors by re-pairing them. To illustrate, for the above example “marcar un número” ‘dial a number’, the verb “entregar” (from “entregar una propuesta” ‘submit a proposal’) was used as a distractor. This set of practice materials was therefore identical for both experimental groups. For the rest of the collocations (i.e., the ‘L1-L2 Congruent’ and ‘Experimental Incongruent’ lists), two additional sets were created, one for each of the two experimental learning groups. The sets critically differed in the distractor verbs presented for trials from the ‘Experimental Incongruent’ collocations list but were otherwise identical. For one group, the distractors of ‘Experimental Incongruent’ items were unrelated verbs; we will henceforth refer to this as the ‘Unrelated’ group. The other group of learners saw distractors that would be congruent with the native language (‘L1-Interference’ group). For the above example, “gastar bromas” (‘play jokes’), the ‘Unrelated’ distractor was “ordenar bromas” ‘order’, while the ‘L1-Interference’ distractor was “jugar bromas” ‘play’ (to reiterate, “jugar” was the literal translation of the verb in the English equivalent “play jokes”). An example is shown in Table 4-2. Each of the forty-two collocations was repeated eight times, yielding a total of 112 trials per list, and 336 trials per participant. This is similar to previous studies that presented each collocation nine times (Experiment 1, Chapter 2; Toomer & Elgort, 2019), but allowed for items to be evenly divided into target-first and distractor-first trials.

<table>
<thead>
<tr>
<th>Experimental Incongruent</th>
<th>Congruent</th>
<th>Independent Incongruent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish (L2)</td>
<td>gastar bromas</td>
<td>jugar partidos</td>
</tr>
<tr>
<td>English (L1) equivalent</td>
<td>play jokes</td>
<td>play matches</td>
</tr>
<tr>
<td>Literal L1 translation</td>
<td>‘spend jokes’</td>
<td>‘play matches’</td>
</tr>
</tbody>
</table>
Practice procedure

Participants were capped before beginning the Practice procedure. First, they completed the ‘Independent Incongruent’ block, followed by the recording of the EEG data as participants completed the Experimental block. This order of presentation ensured that the independent block, which was identical for both groups of learners, would not be affected by the experimental manipulation in the second block.

In each Practice trial, participants were presented with two verbs (target and distractor) followed by a noun. First, a fixation cross was presented on the center of the screen, and then one verb was displayed to the left and one to the right of the fixation cross. Responses were made manually after seeing the noun, by pressing a button corresponding to location of the verb (left or right). To control for the order of processing and activation of target and distractor verbs, the two verbs were displayed in sequence (following the design of Experiment 2, Chapter 3). Immediate feedback was provided at the end of the trial by displaying again the correct verb in blue if the response was correct, or in red if the response was incorrect. The timing and sequencing of a trial is illustrated in Figure 4-2.

In critical trials, learners in the “Unrelated” condition saw unrelated distractors, e.g. “ordenar – gastar – broma” (correct response “gastar” ‘spend’), while learners in the “Interference” condition had to correctly discard the L1-equivalent verb, e.g. “jugar – gastar – broma” (i.e., discard “jugar” ‘play’ to choose “gastar” ‘spend’). The order of presentation of target and distractor was counterbalanced, such that “gastar – jugar – bromas” appeared for half of the trials, and “jugar – gastar – bromas” in the other half. Importantly, all verbs were potential candidates, and it was not until the noun was displayed that participants were able to...
select the appropriate verb ("gastar" in this example, based on the familiarized "gastar bromas").

Prior to beginning the experimental trials, participants completed ten practice trials so that they could get used to the task. In preparation for each trial, a “blinking” screen showing two hyphens “- -” was presented for 1100 ms. Participants were instructed to blink during this 1100 ms screen, and to avoid blinking in the 1300 ms second interval during which the words were presented. Words were displayed in font Arial 30.

Figure 4-2: Sample sequence of the Practice procedure.

Testing materials

Three different types of tests were created for the study, serving three different purposes: (1) a Baseline multiple-choice pre-test, which was administered to confirm that participants were not able to recognize the target collocations even when the correct response
was presented to them; (2) a Translation test, in which we assessed the ability to recall the collocation in connection with its meaning in the native language; (3) a contextualized Cloze test, in which we gauged the ability to use the target verbs in context, but in the absence of the L1 equivalents.

**Baseline test**

First, a baseline multiple-choice pre-test was created in order to gauge learners’ potential familiarity with the incongruent collocations at the onset of the study; this would allow us to exclude participants with prior knowledge of the target materials. A sample item is presented in Example 1 below. Each item contained four choices. For the critical incongruent collocations, the choices contained the target verb (e.g., “gastar” ‘spend’), the non-target literal Spanish equivalent (“jugar” ‘play’), and two additional verbs (“hacer” ‘do’, “bromear” ‘joke’) in randomized order.

(1) play jokes - ___________ bromas
   (a) gastar       (b) jugar       (c) bromear       (d) hacer

At the end of the multiple-choice test, participants were asked to provide confidence ratings on a subset of the items for which correct and incorrect responses had been given, including incongruent collocation trials for which the correct response was selected. A scale from 1 to 5 (1 = no knowledge; 5 = certainty in the response) was used.

**Translation Test**

Two different immediate form-recall tests were created to assess learning and retrieval ability. First, an L1-to-L2 translation test was created. For each question in this test, the same English meanings of the Spanish collocations used for the Familiarization phase
were presented (e.g., “play a joke”). This was considered the more stringent of the two post-tests, as it allowed us to assess recall of both meaning and form. Given the need to activate the corresponding meaning in the L1, this test was expected to elicit a greater amount of cross-language interference from the native language.

**Cloze Test**

Additionally, we included a form recall test, in which the nouns in the collocations learned were embedded in a sentence. Participants were asked to use the corresponding verbs that were learned and practiced to complete the gap in each sentence (see Example 2). This test allowed us to assess form retrieval in an L2-only context in which recollection of the meaning of the collocation was not strictly necessary. Rather, the L2-only context created a scenario in which within-language interference from alternative completions was possible, by virtue of the fact that other plausible answers, different from the target verb, were also available to participants based on their previous knowledge. The extent to which a context is high- or low-constraint is well known to influence cloze probability (Kutas & Hillyard, 1984; Federmeier & Kutas, 1999; Molinaro & Carreiras, 2010), affecting the degree to which learners experience within-L2 competition. The sentences created allowed us to test this possibility by manipulating the context: while the sentences for the Experimental Incongruent collocations were high constraining, the contexts for the Independent Incongruent list were less so. In other words, more options were expected to be plausible completions for the Independent Incongruent items, even if participants were explicitly instructed to use the verbs learned.

To validate the materials used for the cloze test, we conducted a norming study with native speakers of Spanish using Amazon’s Mechanical Turk (for additional information on MTurk, see Paolacci & Chandler, 2014; for a validation study, see Sprouse, 2011). We restricted eligible IP addresses to Spanish-speaking countries and collected language background information on all participants. One participant who reported being a non-native
speaker of Spanish was excluded, as well as two participants who failed to finish the survey. We used the responses from 30 native Spanish speakers to calculate Shannon’s entropy scores for verb responses. In this measure, higher values indicate greater entropy. That is, items for which many different responses are provided are associated with greater entropy, i.e., lower certainty in one specific outcome. Although the native speakers in the norming study were not previously familiarized with our target materials, the measure allowed us to capture how constraining the sentences in the Cloze test were. The results confirmed that the sentences for Independent Incongruent items were significantly less constraining (i.e., higher entropy; mean: 3.39, SD: 0.67) than those for Experimental Incongruent items (mean: 2.55, SD: 0.82; $t(24.96) = -2.99, p < .01$). The Cloze test complements the data of the Translation test by providing insight into the role of within-language interference on verb recall across lists, in conditions of minimal L1 interference. The Cloze test materials are available in Appendix E.

**Test administration, scoring and analysis**

The Baseline test was administered in the first session using Qualtrics (Qualtrics, Provo, UT); no feedback was provided. In the second session, the Cloze test was followed by the Translation test. This order was used so that the L1-equivalents would not be shown again until the end of the session. Both tests were completed on a computer by typing the responses. Feedback was provided only after responses in the Translation test, so that erroneous attributions of meanings could be corrected at the end of the session. The Cloze and Translation tests were completed again in the same order when participants returned for the next session. The immediate and delayed post-tests were included in order to assess the effectiveness of practice immediately following the practice, and for comparability with Experiment 1 (Chapter 2). The delay between the immediate and the delayed tests averaged 5.61 days (SD: 0.21); this time lapse did not significantly differ across the Unrelated (mean: 5.75, SD: 1.52) and L1-Interference (mean: 5.48, SD: 0.95) groups ($t(30.99) = -0.69, p = .49$).
To assess the degree of prior familiarity, the accuracy scores and confidence ratings of the Baseline test were used. As described above, participants provided ratings for any correct response in an incongruent trial, using a scale from 1 to 5 (1 = no knowledge; 5 = certainty in the response). Ratings of 1 indicated that the correct option had been selected by chance and were recoded as inaccurate. Participants who indicated previous knowledge of more than one collocation (rating = 5) or substantial familiarity with more than two items (rating ≥ 4) for any of the incongruent lists were excluded from the study. For the remaining subjects, we calculated weighted pre-test familiarity scores for correct responses based on confidence ratings (ratings of 5, 4, 3, and 2 received weights of 0.5, 0.4, 0.3 and 0.2, respectively). Baseline familiarity scores (Unrelated mean: 0.04, SD: 0.1; L1-Interference mean: 0.04, SD: 0.11) revealed no significant differences across groups (t(1296.1) = 0.88, p = .38). Typed responses in the Translation and Cloze post-tests were coded for accuracy. Misspellings were not penalized as long as no more than two phonemes were incorrect, and the response could be interpreted unambiguously. Additionally, in order to conduct an error type analysis, inaccurate responses in the Cloze test were coded for plausibility, i.e., as containing either plausible responses (e.g., “comprar pizza” ‘buy pizza’ instead of “pedir pizza” ‘order pizza’) or implausible responses.

The results reported below were analyzed using mixed-effects logistic regression with the lme4 package (Bates, Maechler, Bolker & Walker, 2015) in R version 3.5.2 (R Core Team, 2018). Given that each participant produced fourteen data points per type of collocation in each test, and because the same participants responded to identical questions, we conducted a combined analysis for both tests, while including Test type as a fixed effect. The fixed effects considered in the analysis included Type of collocation (Congruent, Experimental Incongruent or Independent Incongruent), Group (Unrelated or L1-Interference distractors), Test (Translation, Cloze) and Session (2: immediate tests, 3: delayed tests). Additionally, we also considered the contribution of the individual N400 difference means and its interaction with the factors above. No substantial collinearity was found among the variables considered.
We initially specified a maximally specified random effects structure including all the factors above (Barr, Levy, Scheepers & Tily, 2013). Following initial attempts which led to convergence issues, the random terms were simplified; the reduced structure included by-subject and by-item random intercepts, as well as random by-subject slopes for Type of Collocation and random by-item slopes for Group and N400 means. Model selection was conducted in a forward step procedure, by adding each predictor and their interactions one by one for model comparison; predictors were kept if the model fit was significantly improved (likelihood ratio test, $p < .05$). The reference levels were set to L1-Interference for Group, Congruent for Type of collocations, Cloze for Test type and Session 2. All continuous variables were centered (Baayen et al., 2008). Parameter-specific $p$-values were estimated using the lmerTest package in R (Kuznetsova, Brockhoff & Christensen, 2017). The results and the model output are reported below.

**EEG recording and analysis**

The continuous electroencephalogram (EEG) was recorded from 32 electrodes mounted in an elastic cap (EasyCap; Brain Products, GmbH) and an ActiChamp amplifier (Brain Products, GmbH) with a 24-bit analog to digital conversion (online sampling rate: 500 Hz, 0.1–100 Hz band-pass filter). Electrode impedances were kept below 5 KΩ. During recording, electrodes were referenced to the right mastoid. Grounding electrodes were mounted on the forehead and beneath the right eye. Blinks and eye movements were measured by placing bipolar pairs of vertical electrodes (VEOG) above and below the left eye and lateral electrodes (HEOG) at the outer canthi of both eyes. Preprocessing steps and analyses were performed with MATLAB (R2016a, The Matworks, Inc.) and a combination of scripts and routines implemented in EEGLAB (v. 13.5.4b) and ERPLAB (v. 5.0). Datasets were filtered online with a 25 Hz low pass and 0.1 Hz high pass noncausal IIR Butterworth digital filter. Segments with excessive muscular artifacts on the continuous data were manually rejected. Subsequently, an independent component analysis (ICA) was performed to
extract and reject remaining ocular and muscular artifacts (Jung et al., 2000). Separate within-
group analyses were conducted. The average number of independent components rejected
during ICA averaged 2.17 (max: 4) in the Unrelated group dataset and 2.09 (max: 4) in the
L1-Interference group. Epochs ranging from -200 to 1000 ms after onset of the noun were
extracted from the pre-processed data. All epochs with activity exceeding ±100 µV at any
electrode site were automatically removed using a peak-to-peak moving window (2.29% of
data in Unrelated group; 1.77% in L1-Interference group). Baseline correction was done
relative to pre-stimulus activity. Inaccurate trials were excluded from the analysis (2.31% in
Unrelated group; 4.74% in L1-Interference group). Data from one participant in the L1-
Interference group was excluded due to excessive artifacts, and data from one participant in
the Unrelated group was lost due to experimental error. Data from all remaining participants
(Unrelated N = 22; Interference N = 23) contained at least 30 valid trials per condition.

Based on the predictions, our analysis of the EEG data focused on two main time
windows. In order to analyze differences in the processing of the noun, we analyzed the
canonical N400 window (300–500 ms post-stimulus presentation). Secondly, following
Experiment 2 (Chapter 3), we predicted that a right frontal effect might emerge between 600
and 800 ms for incongruent trials and / or in distractor-last trials. The behavioral data revealed
that average RTs for button presses across conditions ranged from 687 to 759 ms, coinciding
with the expected 600-800 time-window.

Repeated measures ANOVAs were conducted with mean amplitudes as dependent
variables, and Congruency and Order as independent variables. Midline ANOVAs included
Frontality as a predictor (frontal, central, parietal) and were performed on Fz, Cz and Pz.
Lateralized ANOVAs with Frontality and Hemisphere (left, right) predictors were conducted
on F3, F7, F8, FC1, FC2, FC5, FC6, C3, C4, CP1, CP2, CP5, CP6, P3, P4, P7, P8.
Greenhouse–Geisser-corrected values are reported where appropriate. To better characterize
the topography of significant effects, pairwise t-tests were performed on each scalp region
(right frontal: F4, F8, FC2, FC6; left frontal: F3, F7, FC1, FC5; right posterior: CP2, CP6, P4,
P8; left posterior: CP1, CP5, P3, P7). Results of post-hoc paired pairwise t-tests are reported
with FDR-corrected $p$ values. Finally, we performed correlation tests to investigate a potential association between the amplitude of individual subjects’ ERP components and performance in lexical selection as measured accuracy in recall tests. Individual ERP measures for the Congruency effect were calculated as the difference waves of average amplitudes (Incongruent - Congruent); behavioral scores for the congruency-based cost were calculated as the difference between individual averages, taking the accuracy in congruent trials as the individual baseline ($1 - (\text{proportion correct Congruent} - \text{proportion correct Incongruent})$). For example, a learner who had the same proportion of correct responses for both types of collocations (e.g., 0.9 – 0.9) would have a difference score of 0. For ease of interpretability, the resulting score was subtracted from 1 so that positive scores indicate greater accuracy in recall of incongruent collocations, with a value of 1 indicating no difference in accuracy ($1 - 0 = 1$). Correlation tests were conducted between individual subjects’ ERP measures and RTs per condition, as well as between ERPs and accuracy in recall in each test.

**Results**

In what follows, we first present the results of the ERP analysis from the Practice procedure. We then report results of exploratory correlations testing the predicted association between ERP results and recall of Experimental Incongruent trials in the post-tests. Third, we report the analyses testing the effectiveness of the manipulation on recall ability measured in the Translation and Cloze post-tests.

**Results of ERP from Practice**

Here we present the analysis of the EEG data, acquired during the experimental block of the Practice procedure, which contained the trials of Congruent and Experimental Incongruent collocation lists that were subject to the experimental manipulation. Because one main goal of the study was to examine potential differences in components elicited by
different Practice conditions—and given substantial inherent variability across L2 learners—
analyses are conducted separately for each group; as a result, any differences between the
groups are discussed comparatively.

**300-500 ms time window (N400)**

For learners in the Unrelated group, the midline ANOVA revealed main effects of
Congruency \((F(1, 21) = 6.21, p < .05)\) and Frontality \((F(1.26, 26.46) = 3.78, p < .05)\), as well
as a significant Congruency x Frontality interaction \((F(1.26, 26.46) = 2.76, p < .05)\). The
lateralized ANOVA also showed main effects of Congruency \((F(1, 21) = 4.96, p < .05)\) and
Frontality \((F(1.22, 26.84) = 3.77, p < .05)\), as well as two-way interactions for Congruency x
Frontality \((F(1.38, 30.36) = 2.66, p < .05)\), Hemisphere x Frontality \((F(1.36, 38.98) = 5.65, p
< .01)\) and a three-way Order x Hemisphere x Frontality interaction \((F(1.96, 41.16) = 9.96, p
< .001)\). Pairwise comparisons on midline electrodes revealed a significant effect of
Congruency at Cz and Pz, where it was maximal, and no effect at Fz. Main effects for both
Congruency and Order were greater over left posterior electrodes, where all comparisons
were significant with the exception of a trending significant difference between incongruent
distractor-last and congruent target-last trials \((p = .07)\). Significant differences emerged
between congruent distractor-last trials and all other trials in right posterior electrodes. No
effects were found in right frontal electrodes.

For learners in the L1-Interference group, the midline ANOVA showed main effects
of Congruency \((F(1, 22) = 10.37, p < .01)\) and Frontality \((F(1.38, 30.36) = 6.92, p < .01)\).
Similarly, the lateralized ANOVA revealed main effects of Congruency \((F(1, 22) = 9.68, p <
.001)\) and Frontality \((F(1.38, 30.36) = 6.98, p < .01)\), as well as a significant interaction of
Order x Hemisphere x Frontality \((F(2, 44) = 11.59, p < .0001)\). In contrast with the Unrelated
group, follow-up pairwise comparisons on midline electrodes indicated that Congruency-
based differences were maximal at Fz for the L1-Interference group. Congruency produced
significant differences in all electrode regions, with the exception of right posterior
electrodes, where congruent distractor-last and incongruent distractor-last trials were not significantly different ($p = 0.38$). Over left posterior electrodes, only differences between incongruent target-last trials and all other trials emerged. Significant Order effects emerged for congruent trials in right posterior and left frontal regions, as well as a marginal effect ($p = 0.05$) in incongruent trials in left posterior electrodes. The scalp plots in Figure 4-5 show the distribution of the effects of Congruency and Order in each group within 300-500 ms and 600-800 ms. Figures 4-3 and 4-4 present the grand averages for the Unrelated and L1-Interference groups, respectively.

600-800 ms time-window

In the analysis of the Unrelated group dataset, the midline ANOVA revealed significant effects of Congruency ($F(1, 21) = 23.55, p < .0001$) and Frontality ($F(2, 42) = 5.64, p < .01$), and their interaction ($F(2, 42) = 3.47, p < .05$). Similarly, the lateralized ANOVA showed main effects of Congruency ($F(1, 21) = 23.68, p < .0001$) and Frontality ($F(2, 42) = 4.24, p < .05$), and significant two-way interactions of Congruency x Hemisphere ($F(1, 21) = 5.15, p < .05$), Hemisphere x Frontality ($F(2, 42) = 6.68, p < .01$), and a three-way interaction of Order x Hemisphere x Frontality ($F(2, 42) = 60.52, p < .0001$). Pairwise comparisons confirmed a main effect of Congruency in all midline electrodes, maximal at Fz. Over the four regions of electrodes, the effect of Congruency was maximal at right posterior electrodes, and significant in all other regions with the exception of a lack of significant differences between congruent target-last trials and incongruent distractor-last trials over right parietal electrodes. A significant effect of Order also emerged in left frontal and maximally in right parietal electrodes.

For the L1-Interference group, no significant differences were found in the midline ANOVA. The lateralized ANOVA resulted in significant interactions of Hemisphere x Frontality ($F(1.41, 30.93) = 3.74, p < .05$), and Order x Hemisphere x Frontality ($F(2, 44) = 5.25, p < .0001$). Similar to the Unrelated group, pairwise comparisons indicated a significant
effect of Order in posterior electrodes, greatest over the right hemisphere, as well as a less consistent effect in the left frontal region, which reached significance only for congruent trials.

*Correlations between Practice task ERP data and performance in recall post-tests*

Exploratory tests were performed to investigate the potential correlation between the neurophysiological responses elicited during Practice and accuracy in the recall tests that were subsequently completed. Individual ERP indices were calculated for each participant based on the mean amplitude from each condition (Incongruent – Congruent) in the two time windows analyzed (300-500 ms and 600-800 ms, respectively). Positive values in this difference score indicate a more positive N400 in processing the noun of Incongruent collocations relative to Congruent trials, i.e., inhibition of Congruent collocations. Because differences in the N400 time window were greatest at Fz in the L1-Interference group, but maximal at Pz for the Unrelated group, values were calculated for Cz. For the 600-800 ms time window, differences were computed for the mean of the right posterior region. Recall scores for each participant were also calculated as the difference between conditions. Positive values indicate greater ability to recall incongruent collocations (see additional details in the Methodology section). Difference scores for ERPs and behavior were calculated for all participants and correlations were performed on the whole dataset.

For the Translation test, a significant positive correlation was found between difference scores in the N400 time window and accuracy in recall ($r = .31, p < .05$). A marginally significant correlation was found for the Cloze test ($r = .28, p = .07$). The results are illustrated in Figure 4-6. No significant correlations were found between differences in the 600-800 ms time window of ERPs and performance in post-tests.
Figure 4-3: Grand average ERPs of the Unrelated group: (a) Effects of Congruency and Order at Cz; (b) effect of Congruency; and (c) effect of Order at representative electrode locations. Negativity is plotted up.
Figure 4-4: Grand average ERPs of the L1-Interference group: (a) Effects of Congruency and Order at Cz; (b) effect of Congruency; and (c) effect of Order at representative electrode locations. Negativity is plotted up.
Figure 4-5: Scalp topographies of the difference waves showing the effects of Congruency (Incongruent – Congruent) and Order (DT – TD in plausible trials) in the 300-500 and 600-800 ms time windows for each experimental group.

Figure 4-6: Correlations between individual N400 mean differences at Cz and recall accuracy in immediate post-tests.
Discussion of ERP results

We aimed to investigate the neurophysiological signatures associated with different retrieval Practice conditions, by analyzing ERPs that were time-locked to the presentation of nouns, as well as their association with recall in two subsequent post-tests. The results showed that the two experimental groups displayed distinct neurophysiological signatures during completion of the Practice. First, divergent patterns were found for the N400 window across groups. Secondly, significant by-congruency differences emerged after 600 ms only in the Unrelated group. We discuss the results for each time window in turn.

Differences within 300-500 ms

The analysis of the N400 window allowed us to examine differences in the processing of the noun resulting from the distractor verbs that preceded it, while keeping target verbs constant. As in previous experiments that capitalized on collocational priming, we predicted that a known association between the verb and the noun would result in priming during processing of the noun. The effects of priming were expected to be greater when such an association was strong (i.e., in L1-L2 congruent items) in the form of an attenuated N400. For new incongruent collocations, more successful learning should therefore lead to a smaller difference in the N400 of incongruent relative to congruent collocations. At the same time, we hypothesized that if inhibition of L1-L2 congruent collocations played a role in learning (as suggested by the behavioral data from Experiment 1, Chapter 2), we might observe a critical reversal in processing costs: congruent trials would reveal inhibition, potentially observable as an enhanced N400 relative to incongruent trials. In line with these predictions, a significant N400 in the Unrelated group revealed greater facilitation in processing the nouns of congruent vis-à-vis incongruent collocations. Significant differences were also found in the L1-Interference group but, critically, these showed a relative cost for congruent collocations, with a less negative wave for incongruent collocations. The results provide compelling
evidence that practice in selecting the correct verb resulted in inhibition of L1 congruent verbs, when these were plausible alternatives based on the native language (L1-Interference group).

Additionally, we expected that inhibition might play a critical role in the ability to retrieve the verbs of incongruent collocations beyond the Practice task. More specifically, we hypothesized that evidence of inhibition in ERPs during Practice might predict performance in subsequent recall tests. The exploratory correlations between individual ERPs and test scores were also considered important given that individual variability was expected within- and across-groups. Specifically, previous behavioral evidence also suggests that L1 inhibition was present in more successful learners, even across different practice conditions (Experiment 1). One aspect to note is that the N400 for incongruent collocations (Unrelated group) had a centro-parietal distribution, whereas the reversed effect (greater N400 for congruent items) in the L1-Interference group was maximal at frontal electrodes. There is an unsettled debate as to whether the classic centro-parietal N400 and its frontal counterpart (FN400) are associated with differences in functionality, with some researchers arguing that they are equivalent (e.g., Voss & Federmeier, 2012) while others defending that FN400 modulations are attributable to familiarity (e.g., Woodruff, Hayama & Rugg, 2010; Yu & Rugg, 2006). Since our materials cannot directly speak to this issue, it is beyond the scope of this study to adjudicate between the two views. Nevertheless, we note that these differences in the distribution of the effect across the two groups posed a challenge for the correlation analysis including all subjects. For consistency and replicability, we opted to use voltage at Cz, which showed significant differences for both groups.

The results of correlations showed that, indeed, the individual EEG activity during the Practice task was correlated with performance in recall tests conducted later. The association reached significance in the Translation test, in which the ability to resist interference from the L1 was most critical, and trended toward significance in the Cloze test, in which native language equivalents were not directly presented. Altogether, this pattern of results provides compelling evidence that Practice conditions that aim to induce L1-
interference successfully produce inhibition of L1 competitors, and that this is directly associated with higher rates of recall of learned items.

**Differences within 600-800 ms**

For learners in the Unrelated group, modulations in the N400 were followed by later differences in the 600-800 ms window, with a greater negativity in incongruent collocations that was maximal at right frontal electrodes. That is, the negativity was associated with those trials for which uncertainty was higher and for which confidence in recognition would have been lower. This right frontal effect (RFE) has indeed been described in memory studies as an index of generic memory monitoring processes (Hayama, Johnson & Rugg, 2008; Leynes & Kakadia, 2013), and as different from the posterior late positive component (LPC) associated with recollection. The RFE found in this Practice task is consistent with the results from Experiment 2, where a similar paradigm was employed, and in which an RFE was also elicited for all trials requiring enhanced monitoring in response selection. Interestingly, no effect was observed for the L1-Interference group, suggesting that different mechanisms were deployed to avoid incorrect selection of the distractor. Because learners in the L1-Interference group showed a “flipped N400” (i.e., more negative for congruent than for incongruent collocations), this would suggest that the inhibition of distractors acted as a mechanism to reduce interference. That is, inhibition acted in lieu of monitoring processes indexed by the ERPs in the Unrelated group. We therefore propose that the different neurophysiological signatures indicate that different mechanisms were engaged in each group. While Unrelated learners required enhanced monitoring during response selection, learners who practiced selection in conditions of L1 interference exerted inhibition to regulate the influence of the native language. If this interpretation is correct, one would predict that the amplitudes of the RFE should be negatively correlated with the N400 inhibition wave at the individual level. A post-hoc correlation analysis was performed to test this hypothesis, which confirmed that both components were indeed negatively correlated at the individual level ($r =$
An important aspect to note is that, while both components index mechanisms that would have provided effective pathways to perform the task, only the N400 differences associated with inhibition correlated with higher recall ability during subsequent testing.

Finally, Order produced localized differences limited to right parietal electrodes in the 600-800. In both groups of learners, a more positive-going wave was elicited for distractor-last trials. Because distractor-last trials require retrieving a previously presented verb, one might speculate that the sustained positivity may be related to a higher working memory load. This is unlikely, however, because the effect does not share the scalp topography of components associated with variable working memory conditions. Order did not produce any notable differences across groups. Given that the motivation of counterbalancing the order of target and distractor in our design was simply to render the position of the target unpredictable, this is an aspect that may be addressed in future research.

**Results of Translation and Cloze tests**

Figure 4-7 shows the results of the Translation and the Cloze tests for Session 2 (Immediate tests) and Session 3 (Delayed tests). The model output is presented below in Table 4-3, with significant terms and interactions in bold font. The results revealed the expected main effect of Type of collocation such that, relative to Congruent collocations, accuracy in recall was significantly lower for Experimental Incongruent collocations ($\beta$: -1.86, SE: 0.48, $p < .0001$), and for the Independent Incongruent collocations ($\beta$: -2.16, SE: 0.49, $p < .0001$). Comparisons performed by re-referencing Type of collocation revealed no significant differences between the two types of incongruent collocations; however, the interaction of Type of collocation by Test indicated that, relative to the Experimental Incongruent items, recall for Independent Incongruent collocations was significantly lower in the immediate Cloze test ($\beta$: -0.43, SE: 0.2, $p < .05$). The Type of collocation by Session interaction showed that recall was worse for the Experimental Incongruent items in the delayed post-tests relative to the immediate tests.
No significant main effect of Group or interactions were found. However, a significant interaction of the individual N400 difference index and Session emerged, indicating that higher individual N400 means (indicating L1 inhibition) were associated with significantly higher recall in the immediate tests completed after the Practice ($\beta$: 0.17, SE: 0.06, $p < .01$). Finally, a two-way interaction between Session and Test showed that recall was worse in the delayed Cloze ($\beta$: -1.03, SE: 0.25, $p < .0001$); nonetheless, a three-way interaction indicated that there was less of a decline in incongruent items relative to the congruent items in the immediate Cloze test.

**Plausibility-based error analysis of Cloze test**

As indicated above, incorrect responses from the Cloze test were coded for error type based on whether the non-target verbs were plausible or implausible completions for the context provided. The by-group comparisons showed that, for the high-constrain contexts (Experimental Incongruent collocations), responses in the immediate Cloze test were equally implausible for both the Unrelated (mean: 0.05, SD: 0.22) and the L1-Interference group (mean: 0.04, SD: 0.20; ($t$(597.37) = -0.44, $p = .66$); there were also no significant differences in the delayed Cloze test (Unrelated mean: 0.11, SD: 0.31; L1-Interference mean: 0.10, SD: 0.31; $t$(584.79) = -0.4, $p = .69$). For errors in low-constrain contexts (Independent Incongruent list), learners in the Unrelated group produced verbs that were plausible competitors at a marginally significantly higher rate than the L1-Interference group in the immediate Cloze test (Unrelated mean: 0.08, SD: 0.27; L1-Interference mean: 0.13, SD: 0.33; $t$(607.37) = 1.8624, $p = .06$) and at a significantly higher rate in the delayed Cloze test (Unrelated mean: 0.14, SD: 0.35; L1-Interference mean: 0.07, SD: 0.25; $t$(499.33) = -2.84, $p < .01$).
Figure 4-7: Results of recall accuracy in immediate and delayed post-tests across the two experimental groups. Error bars represent 95% confidence intervals.

Table 4-3: Generalized linear mixed effects model output.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.693</td>
<td>0.396</td>
<td>6.780</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Experimental Incongruent</td>
<td>-1.858</td>
<td>0.476</td>
<td>-3.903</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Independent Incongruent</td>
<td>-2.161</td>
<td>0.487</td>
<td>-4.438</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Session</td>
<td>-0.760</td>
<td>0.183</td>
<td>-4.149</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Cloze test</td>
<td>0.008</td>
<td>0.192</td>
<td>0.042</td>
<td>0.97</td>
</tr>
<tr>
<td>N400 difference mean</td>
<td>0.233</td>
<td>0.189</td>
<td>1.232</td>
<td>0.22</td>
</tr>
<tr>
<td>Exp. Incongruent x Session 3</td>
<td>-0.759</td>
<td>0.232</td>
<td>-3.276</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Indep. Incongruent x Session 3</td>
<td>-0.003</td>
<td>0.236</td>
<td>-0.012</td>
<td>0.99</td>
</tr>
<tr>
<td>Exp. Incongruent x Cloze test</td>
<td>0.335</td>
<td>0.237</td>
<td>1.450</td>
<td>0.16</td>
</tr>
<tr>
<td>Indep. Incongruent x Cloze test</td>
<td>-0.091</td>
<td>0.241</td>
<td>-0.377</td>
<td>0.71</td>
</tr>
<tr>
<td>Session 3 x Cloze test</td>
<td>-1.029</td>
<td>0.251</td>
<td>-4.107</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Session 3 x N400 difference mean</td>
<td>-0.173</td>
<td>0.064</td>
<td>-2.712</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
Discussion of recall in post-tests

The results of the behavioral analysis revealed the expected congruency effect in the main interaction of Type of collocation. The data showed a numeric difference between Experimental Incongruent collocations across the two experimental groups, with higher scores for the L1-Interference group. However, despite a numeric trend, there was no main effect of experimental Group nor did it significantly interact with any of the variables considered. Therefore, while the results align with the pattern in Experiment 1 (Chapter 2), the fact that the differences observed did not reach statistical significance in our model gives support to the hypothesis that manual responses (e.g., button presses) engaged memory and control systems to a lesser extent than practice requiring overt vocal responses.

Interestingly, while the effect of Group was non-significant, individual ERPs did significantly correlate with recall rates in the immediate tests. Two aspects are worth noting in this regard. First, the individual ERP indices provided a continuous measure which, in regression models, are not hindered by power loss in the way that categorical variables are. Relatedly, the ERP measure captured individual variability, which also carries theoretical import. While the between-group manipulation was devised with the goal of inducing the observed ERP correlates of L1 inhibition, earlier behavioral data from RTs in Experiment 1 (Chapter 2) had already shown that slower responses for congruent collocations were associated with recall in subsequent testing. More importantly, those measures indicated that, although inhibition in RTs was mostly present in the L1-Interference group, some individuals successfully learned incongruent collocations showed inhibition, even if they were trained in the less enabling condition with unrelated distractors. Therefore, individual variability is to be
expected, and the ERP data are additional evidence of a direct association between cognitive processes measured at the individual level during Practice and performance in testing.

Finally, the use of the Cloze test and the manipulation therein provided insight into the effectiveness of retrieval when no equivalent from the L1 was provided. First, while no main effect of Test was found, an interaction with session revealed that recall was lower for the delayed Cloze test. This is not surprising, given the fact that selecting the correct verb in this test required remembering both the target meaning and form, whereas in the Translation test the exact target equivalent was provided. That is, it is quite plausible that learners experienced some difficulty remembering the particular action to be expressed in the verb, from among potential completions. Secondly, the role of competition (within the L2) was more directly examined through the manipulation affecting how constraining the context was. The logistic regression analysis showed that recall was in fact lower for the list of Independent Incongruent collocations, which were presented in less constraining contexts. This is preliminary evidence of the role of context constrain, as expected from the results of the norming study, which showed greater entropy in native speakers’ completions of those same items. While this manipulation seemed to be effective, no interaction with Group was found. Therefore, the accuracy rates cannot speak to whether training in suppressing L1-Interference conditions might have afforded any additional benefits. However, the error type analysis conducted did provide insight into this issue, as it showed that learners in the L1-Interference condition made fewer errors containing non-target verbs, which suggests a lower incidence of within-language competition in that group. The results therefore give initial support to the proposal that there may be some benefits from training in specific conditions that generalize to retrieval of other items, although this aspect will need to be further investigated in future research.
Discussion

By measuring EEG and behavior, the goal of Experiment 3 was to investigate the cognitive mechanisms engaged by different practice conditions in real time, and to examine the association between brain potentials and subsequent performance in testing. In other words, the present experiment sought to identify the neural correlates of the desirable difficulty identified in Experiment 1 (Chapter 2) during practice in L2 learners, and the association with learning outcomes.

The results of Experiment 3 revealed that different conditions of retrieval practice, which differed solely in the type of distractors included, elicited different patterns of neurophysiological activity in each group of participants. At the group level, participants in a control condition with unrelated distractors (the ‘Unrelated’ group) showed an N400 effect for incongruent collocations, indicating a cost in processing incongruent vis-à-vis congruent items. Additionally, they presented a right frontal effect (RFE) for trials with incongruent collocations, associated with increased monitoring during response selection. The RFE was not predictive, however, of future ability to recall the learned items in subsequent immediate and delayed tests. A very different pattern was found in the group of learners who practiced under “desirably difficult” conditions, i.e., with distractors that were acceptable based on the native language (‘L1-Interference’ group). The group with L1-interference distractors presented a “reversed N400”, i.e., a greater cost in processing congruent relative to incongruent collocations. Because in the experimental manipulation the verbs of congruent collocations were the distractors to be rejected during incongruent-collocation trials, the N400 for congruent trials provide compelling evidence that distractors were inhibited.

Critically, the degree of inhibition in this reversed N400 was correlated with behavioral measures of learning. The ERP data were compared with the results of recall post-tests, to test the hypothesis that L1 inhibition is associated with L2 attainment. The results confirmed that the relative magnitude of N400 components at the individual level was a significant predictor of performance in the post-test completed immediately after training.
That is, a greater cost in processing *congruent* collocations in ERPs was associated with more accurate recall of incongruent collocations in testing. Therefore, the results of the experiment identified inhibition, as indexed by the reversed N400, as a neurophysiological correlate associated with the practice condition that posed a desirable difficulty for learning of L2 collocations.

**Divergence in ERP components indicate alternative strategies to practice and learning**

Differences in the components elicited at the group level suggested that practice conditions had an impact on the way selection of correct responses was approached during the task. While participants in the Unrelated group presented a clear RFE associated with monitoring response selection, those in the L1-Interference group showed no such effect, but a reversed N400. A subtle manipulation (presenting different distractors to each group) appeared to have deeply influenced participants’ strategy and the cognitive mechanisms engaged to complete the task. A negative correlation between the RFE and the reversed N400 at the individual level confirmed that the trend in participants to show one component or the other. This indicates that individuals relied predominantly on one of two strategies: either monitoring closely for the response to be selected, or inhibiting the strongly competing distractors to reduce interference during response selection. This finding illustrates how purportedly small changes in the input provided to learners during practice may result in different cognitive paths being taken to accomplish the same task, each with a different level of effectiveness and unequal consequences on measurable learning outcomes.

**Indirect evidence for the effect of modality in response selection**

A secondary goal of Experiment 3 was to test whether the group differences produced by a comparable manipulation in Experiment 1, in which response selection was made vocally, would be replicated through manual response selection. The results revealed that the
experimental between-group manipulation was still effective in engaging inhibitory mechanisms as shown by the ERP data. However, at the group level, there was a lack of significant differences in post-test performance. That is, the by-group comparison of behavioral recall alone would have suggested that the manipulation was not effective when responses are made manually; this possibility was ruled out by significant between-group differences in ERPs. The differences between the results of Experiment 1 and the present one, however, lend support to the idea that inhibition training is most effective when exerted directly on oral production, but less so when suppression of competitors is more indirect and mediated by manual responses. Unlike vocal gestures, button presses are not directly linked to linguistic representations, but to somewhat arbitrary task-related mappings, i.e., in this case target responses were identified according to their position on the screen (left or right). If correct, the hypothesis that manual responses rely on linguistic representations indirectly would have practical implications for training in face-to-face and online language instructions. However, because the present experiment is a conceptual replication of Experiment 1 (with an analogous design but a different set of materials) the comparison between the results of both tests is an indirect one, and this idea will need to be more directly replicated in future studies.

Consequences of training language control on learning

Two tests were included to assess recall ability when the L1 equivalent is provided (Translation test), and when the test is conducted entirely in the L2 (Cloze test). No differences between the types of tests were found in terms of recall accuracy. However, the sentences of the L2-only Cloze test were manipulated and normed in terms of how constraining they were. Low constrain sentences were expected to induce greater competition from alternative within-L2 competitors, and therefore to potentially elicit a greater number of errors due to responses consisting of plausible non-target verbs. On the other hand, it was hypothesized that the benefits from training in regulating cross-language competition (L1-
Interference group) might carry over to help regulate within-language competition in the L2. The results of the error analysis showed that learners in the Unrelated group produced significantly higher rates of plausible non-target responses than the L1-Interference group, suggesting a generalization of the benefits from training regulation ability to suppress within-language competitors.
Conclusions and future directions

The overarching goal of this dissertation was to investigate the association between the conditions of practice during L2 learning and the engagement of cognitive mechanisms, as well as their impact on recall.

Chapter 1 (“Introduction”) reviewed prior research on the processing and learning of L2 multiword units, including relevant findings on language non-selective lexical access, the role of the native language during learning of L2 collocations, and the role of cognitive control on resolving interference and competition in retrieval of L2 multiword units. Next, drawing on the literature on “desirable difficulties” in learning, as well as on recent calls to apply this approach to L2 learning (Bjork & Kroll, 2015; Suzuki, Nakata & DeKeyser, 2019), it was proposed that native language regulation and interference inhibition may be viewed as desirable components in the learning of L2 collocations.

Chapters 2-4 reported on three experiments which, together, tested the hypothesis that successful retrieval of L2 collocations hinges on the ability to regulate the L1, and that this skill may be developed in the appropriate practice conditions. The experiments investigated (a) whether certain input conditions aid to develop the ability to regulate interference from the L1 (Experiment 1, Chapter 2); (b) what brain event-related potentials are elicited during multiword lexical selection, to identify the mechanisms involved in retrieval of collocations in both the L1 and the L2 (Experiment 2, Chapter 3); the connection between practice conditions, the brain activity elicited in each condition, and later behavioral performance (Experiment 3, Chapter 4).
Summary of results

Summary of Experiment 1

Experiment 1 tested Spanish learners of English on a new paradigm that aimed to induce interference from the native language during lexical selection in a second language, as a way to train regulation of the dominant language. The results from immediate and delayed L1-to-L2 translation tests showed that recall rates were significantly higher in the group of learners that practiced in conditions of L1-interference. Faster RTs from vocal responses during practice showed more efficient lexical selection in those same learners. Additionally, RTs revealed that the more successful learners in both groups incurred a cost in accessing verb choices congruent with the native language, a finding that is consistent with an inhibitory account.

Summary of Experiment 2

In Experiment 2, event related potentials (ERP) were used to investigate the mechanisms underlying lexical retrieval and selection in conditions of high and low conflict, similar to those elicited in the first experiment. Native and non-native English speakers were asked to select familiarized target verb–noun sequences (e.g., *eat breakfast*) when given two choices. Trials were either low-conflict, with only one plausible candidate (e.g., *eat – shoot – breakfast*) or high-conflict, with two plausible verbs (e.g., *eat – skip – breakfast*). While costs in response selection were modulated by plausibility and order in native speakers, only consistent effects due to plausibility were found in non-native speakers. Additionally, brain activity was only significantly correlated with performance in native speakers. The results suggest largely similar basic mechanisms, but also that different resources and strategies are engaged by non-native speakers when resolving conflict in the weaker language, with a greater focus on individual words than on multiword units.
Summary of Experiment 3

Taking direction from the first two studies, Experiment 3 aimed to directly examine the association between retrieval practice conditions, neurophysiological activity, and learning outcomes. English-speaking learners of Spanish completed a learning, practice and testing paradigm similar to the one used in the first experiment. The results revealed different neurophysiological signatures for participants in each of the two practice conditions. Learners who saw unrelated distractors presented a greater N400 for incongruent relative to congruent collocations, as well as a response-selection related right frontal effect (RFE). However, learners who saw L1-related distractors presented no RFE, but revealed instead a reversed N400 effect – i.e., a more negative peak for congruent than for incongruent collocations –, indicating L1 inhibition. The analyses revealed a direct association between ERPs during practice and learners’ ability to recall collocations, such that greater inhibition in congruent ERP trials predicted higher accuracy in the immediate post-tests.

Implications for second language learning

The relevance of the desirable difficulties framework

The results of Experiments 1 and 3 provided converging evidence from behavior and ERPs showing that, in cases of competition between the dominant L1 and the weaker L2, resulting in a cost in access to the L1 may be necessary to alleviate costs in accessing the L2. This pattern of results is in line with the rationale of the “desirable difficulties” approach, in which particular conditions may be found to pose an initial relative burden during learning, but which are in fact a consequence of engaging the cognitive mechanisms leading to long-term retention (Bjork, 2018). The experiments reported here tested, and provided support to, the hypothesis that L1 inhibition is a desirable difficulty in acquiring L1-L2 incongruent collocations.
The research into L2 acquisition will benefit from future studies framed within a desirable difficulties approach that seek to identify and explore more effective training conditions that lead to stable learning outcomes (for a recent proposal along these lines, see Suzuki, Nakata & DeKeyser, 2019). An important aspect to note, however, is that this type of desirable difficulty (inhibition of L1 counterparts) and the mode of training (inducing interference) should be regarded as specific to the case at hand, and the form and type of desirable difficulties in other scenarios will be contingent on the mechanisms involved (for an overview see Bjork, 1994). These will often by, by definition, counter-intuitive – either because they do not immediately lead to short-term gains, they result in an initial cost or produce in some way a non-linear trajectory towards ultimate L2 development. Such examples already exist in the literature on L2 learning, even if the term “desirable difficulties” was not employed. For example, in the L2 acquisition of Spanish subjunctive verb mood, a notorious challenge is to get learners to notice when the subjunctive should be used. Part of the problem stems from the fact that, in regular verbs, the subjunctive is marked by a vowel change at the end (trabaj{a, e}) or in the middle of the verb (trabaj{a, e}mos). Because learners fail to notice the change in the verb declension, they also fail to notice the contexts in which the subjunctive appears (Buckwalter, 2001; Fernández, 2008; Leow, 1993). A counter-intuitive solution to this obstacle rests in relying on irregular verbs. Although learning the irregular verb forms is in itself an added difficulty, these are more easily noticed and therefore may facilitate the development of an association between form and function in a given context (Collentine, 1997; Gudmestad, 2006; for a review, see Collentine, 2010). While previous studies did not frame this approach as a “desirable difficulty”, it certainly fits the previously discussed criteria. The prospect of ameliorating instruction by understanding and addressing specific cognitive challenges of L2 acquisition opens up exciting avenues of research. In the case of Spanish, these might focus on aspects such as the acquisition of the subjunctive, past verb tenses, or various L1-L2 incongruent constructions (e.g., verbs of “becoming”, Bybee & Eddington, 2006).
Implications for language instruction

One goal of the research contained in this dissertation is to produce findings that will inform language learning practice. In fact, some of the training procedures employed can be easily translated to practice in the classroom and to the design of instruction materials. First, the design of the practice procedures described in Experiments 1 and 3 can be implemented on computer-based practice with minimal modifications. Similarly, the same designs can be implemented in face-to-face activities that are everyday practice in the classroom, e.g., providing students with index cards that contain the options (verbs) and noun to be read out loud for a partner to select a response, and to then provide immediate oral feedback. On the other hand, additional classroom-based research will be needed to confirm the effectiveness of training conditions, especially if practice in the classroom requires additional modifications, such as completing a lower number of repetitions. Therefore, a logical continuation of the laboratory-based work presented here will be to investigate its implementation in second / foreign language courses.

Another potential extension of this work concerns the adaption of the materials to difference audiences, and in particular to groups of learners with different language backgrounds. The materials used in the three experiments conducted were designed with populations of English and / or Spanish speakers in mind. Nonetheless, the same logic (i.e., engaging language regulation through distractors that are L1 equivalents) can be applied to materials adapted to speakers of other languages. In fact, it is already common practice among textbook publishers to customize part of the content as appropriate to speakers of particular language groups, in order to address specific challenges that differ for each language pair (see e.g., Cunningham, Redston & Bell, 2013; Edwards, Gairns, Redman & Rimmer, 2016, for adapted versions). In this vein, hybrid and online L2 instruction offer the possibility to go one step further, and to tailor practice materials for students based on their L1, even when they are part of a group of individuals of various linguistic backgrounds.
The role of production

The discussion of Experiment 3 highlighted some important differences between the results obtained through the paradigm of retrieval practice with manual responses, and the effect of vocal responses in Experiment 1. Most notably, while vocal responses resulted in significant behavioral differences by group, manual selection did not. As noted, the comparison between both response modalities is only indirect given that, while both experiments were designed based on the same methodology, the materials in the experiments were not identical.

A connection can be drawn between the results in this dissertation and previous findings within the corrective feedback literature. In particular, the effect of production on retrieval ability suggested in our data is in line with the finding that corrective feedback is most conducive to learning when learners engage in production (Havranek, 2002; Lyster, 2004; McDonough, 2005). Extensive research has investigated the question of what type of corrective feedback is more effective. In particular, previous studies have compared the effectiveness of feedback that asks students to rephrase their output (‘modified output’) with cases in which production of the correct form is not required. For example, in some cases feedback is provided through so-called “recasts”, in which the non-target output produced by the learner is reworded (or recast) by an interlocutor (usually the instructor). In this case, learners are less likely to repeat the modified utterance or self-correct (Anton, 1999; Linnell, 1995; Lyster, 1998, 2004; Lyster & Ranta, 1997; Oliver, 2000; Panova & Lyster, 2002). However, recasts have been found to be most effective when the learner is also asked to produce the modified, correct utterance (McDonough, 2005; McDonough & Mackey, 2006). More generally, corrective feedback in which the learner is pushed to produce modified output (e.g., clarification questions) has been found to be more directly linked to L2 development (Havranek, 2002; Li, Zhu & Ellis, 2016; Lyster, 2004; McDonough, 2005). The rationale for the effectiveness of production in error correction may lie in the fact that, once an error is produced, it is critical to re-encode or “overwrite” the non-target retrieval trace
activated while producing it (Izumi, 2003). Overwriting an incorrect retrieval would require executing an equally proactive retrieval episode, in other words, *producing* the correct response, as opposed to simply processing a recast. This body of evidence emphasizes the importance of production in advancing development, especially in cases in which learners have acquired the target L2 representations but fail to produce output that reflects their knowledge. In the context of this dissertation, which proposes that inhibiting non-target representations may be critical for successful retrieval, the findings in the corrective feedback literature underscore the importance of enabling retrieval practice conditions.

The question of whether and when production might contribute to learning resonates within the well-trodden debate between advocates of the Instruction Processing approach (e.g., Vanpatten, 2002; Vanpatten & Cadierno, 1993; Vanpatten, & Price, 2018), who tend to downplay the role of production, and proponents of interaction- and output-oriented approaches (e.g., Gass, 2003; Long, 1996; Swain 1995, 2000). While the debate has calmed down considerably, it may be as relevant as ever in the current context, in which online and hybrid instruction are becoming ubiquitous, even if they often tend to rely on processing more than on production. In the current emergent scenario of increasingly computerized L2 instruction, a wealth of new tools is becoming available to language learners, including telecollaboration with conversation partners (Anderson & Corbett, 2013; van der Zwaard & Bannink, 2014, 2018), augmented reality-based immersion (Godwin-Jones, 2016) or computer-based vocabulary spaced repetition (Chukharev-Hudilainen & Klepikova, 2016). However, many of the more popular tools among instructors and language learners still rely on practice that does not allow or fails to encourage production, e.g., Gimkit (Gimkit, Inc.), Kahoot! (Kahoot, Inc.) Quizizz (Zendesk, Inc.) or Quizlet (Quizlet, Inc.). The implications of output for learning, in particular in what concerns CALL and new technology-based approaches to learning, open new avenues of research connected to the future of language instruction.
Beyond multiword units: Generalization in L2 learning

The results in this dissertation are in line with previous research that has evidenced that learning L1-L2 incongruent multiword units requires a considerable amount of time and effort. However, limited classroom time often does not allow for a sufficient number of units to be explicitly taught, practiced and learned. This leads to questions about what can be done to make the learning of multiword units more efficient.

Given the time constraints most L2 classroom learners face, an approach to learning that allows students to productively increase their linguistic repertoire based on limited input is highly desirable. The basis of productivity in language is linguistic generalization (Bybee, 2010; Clausner & Croft, 1997), i.e. the production and comprehension of novel utterances based on previously known ones. That is, rather than having to learn every idiomatic combination one-by-one, linguistic generalization allows for similar novel utterances to be produced based on a small number of successfully learned examples. Generalization is achieved through partly fixed, partly open linguistic representations, termed constructions (e.g., Goldberg, 1995, 2003, 2019). For instance, learning a phrase such as “ponerse nervioso” (‘get nervous’) might serve as the gateway to produce other similar related phrases, e.g., “ponerse contento” ‘get happy’, “ponerse nostálgico” ‘get nostalgic’, etc., through generalization based on a more abstract “ponerse + [ADJECTIVE]” construction. In fact, evidence shows that native speakers can generalize with ease, and that after one or more prototypical examples are known, they may become the basis for generalization when using a specific construction, such as [“ponerse” + adjective] (Bybee & Eddington, 2006).

A promising direction in future research is therefore one that would focus on constructions, regarded as a vehicle for linguistic generalization. However, generalization is believed to be largely based on similarity (Bybee, 2003; Fisher, Godwin, Matlen, 2015), and languages such as English and Spanish do not completely overlap; for instance, adjectives that are used with the same verb in one language may be used in combination with different verbs in the other language. To illustrate, the verb “become” can be used both in “become
blind” and “become nervous” in English, but different verbs must be used in Spanish (i.e., “quedarse ciego” ‘become blind’ and “ponerse nervioso” ‘become nervous’). Thus, learners often need to acquire entirely new generalizations in their L2, and use language productively in ways that differ from their native language. Nonetheless, most work on generalization has been conducted employing artificial languages (Perek & Goldberg, 2015; Tao & Williams, 2019; Vujović, Ramscar & Wonnacott, 2019; Wonnacott, Newport & Tanenhaus, 2008).

Given the promise of a learning approach based on generalization, some important questions arise: first, are L2 speakers able to categorize and generalize conventional expressions in a similar way as native speakers, as well as overcome constructions from their L1? And if so, what type of input will enable learners to acquire and generalize L2-specific constructions? That is, should input showcase prototypical high-frequency exemplars or reflect the statistical regularities and variability found in native speech? Future research may address these questions by establishing the cognitive mechanisms and input conditions that may enable learners to successfully generalize in a second language.
References


_Psychological Science in the Public Interest, 10_, 89-129.
doi:10.1177/1529100610387084


Collentine, J. (2010). The acquisition and teaching of the Spanish subjunctive: An update on current findings. *Hispania, 39*-51


Davies, Mark. (2016-) *Corpus del Español: Two billion words, 21 countries*. http://www.corpusdelespanol.org/web-dial/


https://doi.org/10.3758/s13423-015-0816-5


doi:10.1017/S136672891000012X


Acquisition of Spanish and Portuguese as First and Second Languages (pp. 170-184).

Cascadilla Proceedings Project Somerville, MA


Saito, K. (2019). Multi-or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning, 70*.


https://doi.org/10.1080/09658211.2018.1510966
Appendix A

Individual measures of linguistic and cognitive ability (Experiment 1)

Participants were administered a number of measures to assess language proficiency in the L1 and L2, as well as tasks designed to ascertain individual differences in cognitive control. These are described below.

1. Measures of language proficiency and language ability

1.1. Language experience and proficiency questionnaire

To assess linguistic proficiency and background in the L1 and the L2, participants completed an abridged version of the LEAP-Q (Marian, Blumenfeld, & Kaushanskaya, 2007). Through the LEAP-Q, data about an individual’s L2 learning background were collected (Onset of Acquisition, weekly exposure, immersion) and other measures relevant to assess learning ability (age, level of education).

1.2. L2 proficiency

General English proficiency was measured by administering an abridged version of the Michigan English Language Institute College English Test (MELICET). Participants also completed a vocabulary test to measure general lexical knowledge, and a multiple-choice test to assess knowledge of the individual words employed in the experiment. High scores in the vocabulary tests demonstrated sufficient knowledge of single words (Unrelated mean: 92%; SD: 4.9; L1-interference mean: 91%; SD: 6.3), with no significant differences between the two groups ($t(35.7) = -0.31, p = 0.76$).

1.3. Phonological short-term memory

Previous research has shown that phonological short-term memory (PSTM) predicts vocabulary learning (Baddeley, Papagno, & Vallar, 1988; Martin & Ellis, 2012). Given that learning multi-word units requires the ability to hold sequences longer than individual words
in memory, PSTM was hypothesized to be of particular importance. A nonword repetition task was used to measure PSTM. Participants heard lists of one-syllable nonwords and repeated each list out loud. The lists employed here were adapted from Martin & Ellis (2012) to conform to Spanish phonotactics. Four lists of three, four, five or six nonwords were presented in ascending order (sample stimuli are provided below in Table 1). Participants’ responses were recorded using a Zoom 4HN Pro digital recorder, and were scored following the criteria described in Gathercole, Pickering, Hall & Peaker. (2001, p. 15).

2. Measures of Cognitive control

To ensure that the two groups were comparable in terms of their cognitive control abilities, participants completed the AX-Continuous Performance Task (AX-CPT), the Flanker task, a Spanish working memory test.

2.1. AX-CPT

The AX-CPT has been used to measure individual styles of cognitive control. Participants saw sequences of five letters, the first and last of which were displayed in red font, and were asked to respond “yes” if the sequence started with the letter “A” and ended with an “X” (A – X). In all other conditions, they were instructed to respond “no.”

The Dual Mechanisms of Control model (Braver, Gray & Burgess, 2007) considers the contributions of proactive and reactive control subcomponents. Because in some trials participants can anticipate a “no” response (B – X) while in others they have to react rapidly (A – Y), an individual’s Behavioral Shift Index (BSI) provides a relative measure of proactive/reactive control, with higher values indicating higher proactive control. Composite BSI scores were calculated following the methods described in Braver, Paxton, Locke and Barch (2009).
2.2. Flanker Task

The Flanker task provided an additional measure of executive control. A large body of studies have used the Flanker task to measure the association between language background and cognitive control (e.g., Emmorey, Luk, Pyers & Bialystok, 2008). In this task, participants responded to the direction of a central arrow in trials in which it is flanked by arrows pointing in the same direction (congruent trials), or the opposite direction (incongruent trials). The Flanker effect for each group was calculated by substracting reaction times in congruent trials from incongruent trials.

2.3. Reading Span Working Memory test

Participants were administered a Spanish version of the Reading Span Task, based on the Spanish adaptation of the original task in Daneman and Carpenter (1980) (Elosúa, Gutiérrez, García Madruga, Luque & Gárate, 1996). To ensure that sentences were processed as a whole, a manipulation was added in which participants were asked to judge whether each sentence made sense or not. As in Harrington and Sawyer’s task (1992), half of the sentences were made ungrammatical by mixing up the order of the words.

Table A1. Sample stimuli of nonword repetition task

<table>
<thead>
<tr>
<th>List length</th>
<th>3 items</th>
<th>4 items</th>
<th>5 items</th>
<th>6 items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lib</td>
<td>dren</td>
<td>yir</td>
<td>cak</td>
</tr>
<tr>
<td></td>
<td>chol</td>
<td>glach</td>
<td>gab</td>
<td>nej</td>
</tr>
<tr>
<td>trum</td>
<td>nit</td>
<td>brok</td>
<td>yat</td>
<td></td>
</tr>
<tr>
<td></td>
<td>lon</td>
<td>tep</td>
<td>mur</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chom</td>
<td>fram</td>
<td></td>
<td>miz</td>
</tr>
</tbody>
</table>
Appendix A References


Appendix B

List of stimuli (Experiment 1)

The table below presents the three lists of collocations that were studied and practiced by learners. Below each English collocation, its idiomatic Spanish translation is provided. The literal English translations for the verbs of L1-L2 incongruent Spanish collocations are provided to the right of each verb.

<table>
<thead>
<tr>
<th>Incongruent collocations</th>
<th>Collocations with L1-related verb</th>
<th>Collocations with semantically-related verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb (det.) noun</td>
<td>verb (det.) noun</td>
<td>verb (det.) noun</td>
</tr>
<tr>
<td>launder</td>
<td>money</td>
<td>clean one's hands</td>
</tr>
<tr>
<td>blanquear ‘whiten’</td>
<td>dinero</td>
<td>limpiar manos</td>
</tr>
<tr>
<td>pack one’s bags</td>
<td>make a cake</td>
<td>ready the room</td>
</tr>
<tr>
<td>hacer ‘make’ maletas</td>
<td>hacer tarta</td>
<td>preparar habitación</td>
</tr>
<tr>
<td>run a business</td>
<td>carry one’s name</td>
<td>walk a street</td>
</tr>
<tr>
<td>llevar ‘carry’ negocio</td>
<td>llevar nombre</td>
<td>caminar calle</td>
</tr>
<tr>
<td>shoot a movie</td>
<td>roll a ball</td>
<td>fire a gun</td>
</tr>
<tr>
<td>rodar ‘roll’ pelicula</td>
<td>rodar pelota</td>
<td>disparar pistola</td>
</tr>
<tr>
<td>file a complaint</td>
<td>put an end</td>
<td>arrange a meeting</td>
</tr>
<tr>
<td>poner ‘put’ queja</td>
<td>poner fin</td>
<td>organizar reunión</td>
</tr>
<tr>
<td>perform a song</td>
<td>touch one’s hair</td>
<td>show pictures</td>
</tr>
<tr>
<td>Spaniard</td>
<td>Catalan</td>
<td>English</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>tocar ‘touch’</td>
<td>canción</td>
<td>touch</td>
</tr>
<tr>
<td>raise</td>
<td>doubts</td>
<td>doubts</td>
</tr>
<tr>
<td>despertar ‘wake’</td>
<td>dudas</td>
<td>wake</td>
</tr>
<tr>
<td>meet</td>
<td>a</td>
<td>target</td>
</tr>
<tr>
<td>cumplir ‘accomplish’</td>
<td>objetivo</td>
<td>accomplish</td>
</tr>
<tr>
<td>take</td>
<td>a</td>
<td>walk</td>
</tr>
<tr>
<td>dar ‘give’</td>
<td>paseo</td>
<td>give</td>
</tr>
<tr>
<td>miss</td>
<td>one’s</td>
<td>flight</td>
</tr>
<tr>
<td>perder ‘lose’</td>
<td>vuelo</td>
<td>lose</td>
</tr>
<tr>
<td>buy</td>
<td>tiempo</td>
<td>win</td>
</tr>
<tr>
<td>ganar ‘win’</td>
<td>tiempo</td>
<td>war</td>
</tr>
<tr>
<td>land</td>
<td>a</td>
<td>job</td>
</tr>
<tr>
<td>conseguir ‘achieve’</td>
<td>trabajo</td>
<td>a</td>
</tr>
<tr>
<td>stuff</td>
<td>one’s</td>
<td>mouth</td>
</tr>
<tr>
<td>llenar ‘fill’</td>
<td>boca</td>
<td>a</td>
</tr>
<tr>
<td>blow</td>
<td>the</td>
<td>bridge</td>
</tr>
<tr>
<td>volar ‘fly’</td>
<td>puente</td>
<td>a</td>
</tr>
<tr>
<td>play</td>
<td>a</td>
<td>joke</td>
</tr>
<tr>
<td>gastar ‘spend’</td>
<td>broma</td>
<td>a</td>
</tr>
</tbody>
</table>

**English Phrases:***
- increase the age
- join one's friends
- catch some breath
- find the truth
- pay the bill
- reach a height
- load a truck
- break one's heart
- relax the mind

**Spanish Phrases:**
- enseñar imágenes
- aumentar edad
- unirse amigos
- coger aliento
- encontrar verdad
- pagar factura
- alcanzar altura
- cargar camión
- romper corazón
- relajar mente
Appendix C

Results of statistical analyses (Experiment 1)

1. Accuracy in recall tests

Table C1. Model output for Recall Accuracy in Immediate and Delayed Tests

|                          | Estimate | Std. Error | |z|   | Pr(>|z|) |
|--------------------------|----------|------------|-----|----|---------|
| (Intercept)              | 2.92     | 0.51       | 5.78|   | <0.0001 |
| Condition Interference   | 0.8      | 0.45       | 1.78|   | 0.07    |
| Type Incong              | -2.03    | 0.58       | -3.49|  | <0.001  |
| Type Semant              | -0.34    | 0.59       | -0.57|  | 0.57    |
| TestDelayed1             | -0.62    | 0.23       | -2.7 |  | <0.01   |
| TestImmediate2           | 1.66     | 0.3        | 5.5  |  | <0.0001 |
| TestDelayed2             | 1.66     | 0.3        | 5.5  |  | <0.0001 |
| TestDelayed3             | 0.66     | 0.26       | 2.57 |  | <0.01   |
| PSTM                     | 0.71     | 0.21       | 3.4  |  | <0.001  |
| Condition Interference * Type Incong | 0.81   | 0.2        | 4.03 |  | <0.0001 |
| Condition Interference * Type Semant | 0.31   | 0.22       | 1.4  |  | 0.16    |
| Type Incong * TestDelayed1 | -0.92 | 0.26       | -3.55|  | <0.001  |
| Type Semant * TestDelayed1 | -0.47 | 0.28       | -1.65|  | 0.10    |
| Type Incong * TestImmediate2 | -0.11 | 0.33       | -0.34|  | 0.73    |
| Type Semant * TestImmediate2 | 0.12  | 0.38       | 0.31 |  | 0.76    |
| Type Incong * TestDelayed2 | -0.11 | 0.33       | -0.34|  | 0.73    |
| Type Semant * TestDelayed2 | 0.12    | 0.38       | 0.31 |  | 0.76    |
| Type Incong * TestDelayed3 | -0.9    | 0.28       | -3.16|  | <0.01   |
| Type Semant * TestDelayed3 | -0.15   | 0.32       | -0.46|  | 0.64    |
| Condition Interference * TestDelayed1 | -0.58 | 0.21       | -2.76|  | <0.01   |
| Condition Interference * TestImmediate2 | -0.3    | 0.27       | -1.09|  | 0.27    |
| Condition Interference * TestDelayed2 | -0.3    | 0.27       | -1.09|  | 0.27    |
| Condition Interference * TestDelayed3 | -0.49  | 0.23       | -2.18|  | <0.05   |

Dedicated Analyses on Delayed Tests 1 and 3

Delayed Test 1

|                          | Estimate | Std. Error | |z|   | Pr(>|z|) |
|--------------------------|----------|------------|-----|----|---------|
| (Intercept)              | 2        | 0.43       | 4.7  |  | <0.001  |
| Condition Interference   | 0.58     | 0.36       | 1.63 |  | 0.10    |
| Type Incong              | -2.4     | 0.48       | -4.97|  | <0.001  |
| Type Semant Rel          | -0.86    | 0.49       | -1.77|  | 0.08    |
| PSTM                     | 0.57     | 0.18       | 3.15 |  | <0.001  |
### 2. Growth Curve Analysis of Reaction Time for verb selection in incongruent collocations

Table C2. GCA output for RTs in incongruent collocations

| Practice session 1 | Estimate | Std. Error | $|t|$ | Pr($>|t|)$ |
|-------------------|----------|------------|------|-----------|
| (Intercept)       | -0.04    | 0.06       | -0.59| 0.56      |
| ot1               | -0.88    | 0.29       | -2.98| <0.01     |
| ot2               | 0.64     | 0.22       | 2.94 | <0.01     |
| Condition Interference | -0.06 | 0.05       | -1.07| 0.29      |
| PSTM              | -0.05    | 0.02       | -2.90| <0.01     |
| ot1 * Condition Interference | -1.09 | 0.42       | -2.59| <0.01     |

| Practice Session 2 | Estimate | Std. Error | $|t|$ | Pr($>|t|)$ |
|-------------------|----------|------------|------|-----------|
| (Intercept)       | -0.15    | 0.05       | -3.12| <0.01     |
| ot1               | -0.71    | 0.27       | -2.63| <0.01     |
| ot2               | 0.43     | 0.17       | 2.47 | <0.05     |

| Practice Session 3 | Estimate | Std. Error | $|t|$ | Pr($>|t|)$ |
|-------------------|----------|------------|------|-----------|
| (Intercept)       | -0.07    | 0.06       | -1.07| 0.28      |
| ot1               | -0.38    | 0.22       | -1.73| 0.08      |
| ot2               | 0.27     | 0.17       | 1.6  | 0.11      |
| Condition Interference | -0.1  | 0.04       | -2.23| <0.05     |
3. Analysis of RT for verb selection across collocation types

Table C3. Results of the RT mixed-effects regression analysis of verb selection for collocation types.

|                                | Estimate | Std. Error | |t|  | Pr(>|t|) |
|--------------------------------|----------|------------|------|---|----------|
| Practice session 1             |          |            |      |   |          |
| (Intercept)                    | 0.27     | 0.11       | 2.52 | 0.01 |
| PSTM                           | -0.01    | 0.08       | -0.13| 0.9 |
| Type Semantic                  | 0.01     | 0.09       | 0.08 | 0.94 |
| Type Incongruent               | -0.02    | 0.09       | -0.22| 0.83 |
| Flanker Effect                 | -0.09    | 0.08       | -1.1 | 0.27 |
| PSTM * Type Semantic           | -0.04    | 0.02       | -2.14| <0.05 |
| PSTM * Type Incongruent        | -0.06    | 0.02       | -3.06| <0.01 |
| Type Semantic * Flanker_Effect | -0.05    | 0.02       | -2.54| <0.01 |
| Type Incongruent * Flanker Effect | -0.06  | 0.02      | -2.75| <0.01 |
| Practice session 2             |          |            |      |   |          |
| (Intercept)                    | 0.04     | 0.09       | 0.49 | 0.62 |
| PSTM                           | -0.02    | 0.06       | -0.34| 0.73 |
| Type Semantic                  | -0.03    | 0.09       | -0.36| 0.72 |
| Type Incongruent               | -0.1     | 0.09       | -1.1 | 0.27 |
| PSTM * Type Semantic           | -0.04    | 0.02       | -2.51| <0.05 |
| PSTM * Type Incongruent        | -0.06    | 0.02       | -3.47| <0.001 |
| Practice session 3             |          |            |      |   |          |
| (Intercept)                    | -0.17    | 0.1        | -1.8 | 0.07 |
| Type Semantic                  | -0.01    | 0.07       | -0.14| 0.9 |
| Type Incongruent               | -0.01    | 0.07       | -0.15| 0.88 |
| Condition Interference         | 0.01     | 0.12       | 0.06 | 0.95 |
| PSTM                           | -0.02    | 0.06       | -0.34| 0.73 |
| Condition Interference * Type Semantic | -0.01 | 0.03      | -0.44| 0.66 |
| Condition Interference * Type Incongruent | -0.11 | 0.03    | -3.34| <0.001 |
| PSTM * Type Semantic           | -0.04    | 0.02       | -2.65| <0.01 |
| PSTM * Type Incongruent        | -0.04    | 0.02       | -2.72| <0.01 |
Appendix D

Results of statistical analyses (Experiment 2)

Table D1. Output of generalized linear regression accuracy analysis.

|                      | Estimate | Std. Error | z value | Pr(>|z|) | p value |
|----------------------|----------|------------|---------|----------|---------|
| (Intercept)          | 4.47     | 0.34       | 13.03   | 0.00     | <0.0001 |
| Group Bilingual      | −0.96    | 0.40       | −2.40   | 0.02     | <0.05   |
| Plausible Distractor | −2.37    | 0.31       | −7.66   | 0.00     | <0.0001 |
| Order TD             | −0.65    | 0.32       | −2.04   | 0.04     | <0.05   |
| Plausible Distractor * Order TD | 0.39 | 0.35 | 1.14 | 0.26 | 0.26 |
| Group Bilingual * Plausible Distractor | 1.31 | 0.38 | 3.42 | 0.00 | <0.01 |
| Group Bilingual * Order TD | 1.10 | 0.42 | 2.62 | 0.01 | <0.01 |
| Group Bilingual * Plausible Distractor * Order TD | −1.14 | 0.47 | −2.44 | 0.01 | <0.05 |

Table D2. Output of mixed-effects regression RT analysis.

|               | Estimate | Std. Error | df  | t value | Pr(>|t|) | p value |
|---------------|----------|------------|-----|---------|----------|---------|
| (Intercept)   | −0.28    | 0.12       | 50.6| −2.28   | 0.03     | <0.05   |
| Group Bilingual | 0.39    | 0.16       | 43.02| 2.37   | 0.02     | <0.05   |
| Plausible Distractor | 0.06 | 0.02 | 6224.44 | 3.07 | 0 | <0.01 |
| Order TD      | 0.13     | 0.02       | 6224.69| 6.09   | 0        | <0.0001 |
Appendix E

List of stimuli (Experiment 3)

Collocations list. The table below presents the three lists of collocations that were studied and practiced by learners. Below each Spanish collocation, its idiomatic English translation is provided. The literal English translations for the verbs Spanish collocations are provided to the right of each verb.

<table>
<thead>
<tr>
<th>Experimental Incongruent collocations</th>
<th>Congruent Collocations</th>
<th>Independent Incongruent collocations</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb (det.) noun</td>
<td>verb (det.) noun</td>
<td>verb (det.) noun</td>
</tr>
<tr>
<td>1 pedir ‘ask for’ pizza</td>
<td>ordenar ‘order’ [este] caos</td>
<td>marcar ‘mark’ [un] número</td>
</tr>
<tr>
<td>order pizza</td>
<td>order [this] chaos</td>
<td>dial [a] número</td>
</tr>
<tr>
<td>2 dirigir ‘direct’ [un] negocio</td>
<td>correr ‘run’ millas</td>
<td>hacer ‘make’ [las] maletas</td>
</tr>
<tr>
<td>run [a] business</td>
<td>run miles</td>
<td>pack [the] bags</td>
</tr>
<tr>
<td>3 rodar ‘roll’ escenas</td>
<td>disparar ‘shoot’ [una] pistola</td>
<td>controlar ‘control’ [el] estrés</td>
</tr>
<tr>
<td>shoot scenes</td>
<td>shoot [a] gun</td>
<td>handle [the] stress</td>
</tr>
<tr>
<td>4 blanquear ‘whiten’ dinero</td>
<td>lavar ‘launder’ ropa</td>
<td>volar ‘fly’ puentes</td>
</tr>
<tr>
<td>launder money</td>
<td>launder clothes</td>
<td>blow up bridges</td>
</tr>
<tr>
<td>5 despertar ‘awaken’ dudas</td>
<td>levantar ‘raise’ [la] cabeza</td>
<td>navegar ‘sail’ [la] web</td>
</tr>
<tr>
<td>raise doubts</td>
<td>raise [one’s] head</td>
<td>surf [the] web</td>
</tr>
<tr>
<td>Spaniard</td>
<td>Spanish</td>
<td>English</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>6</td>
<td>poner 'put'</td>
<td>atención</td>
</tr>
<tr>
<td>7</td>
<td>dar 'give'</td>
<td>paseos</td>
</tr>
<tr>
<td>8</td>
<td>perder 'lose'</td>
<td>[un] tren</td>
</tr>
<tr>
<td>9</td>
<td>ganar 'win'</td>
<td>tiempo</td>
</tr>
<tr>
<td>10</td>
<td>gastar 'spend'</td>
<td>bromas</td>
</tr>
<tr>
<td>11</td>
<td>abrir 'open'</td>
<td>[el] camino</td>
</tr>
<tr>
<td>12</td>
<td>montar 'assemble'</td>
<td>fiestas</td>
</tr>
<tr>
<td>13</td>
<td>sacar 'extract'</td>
<td>fuerzas</td>
</tr>
<tr>
<td>14</td>
<td>revelar 'reveal'</td>
<td>fotos</td>
</tr>
</tbody>
</table>
Cloze test materials

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ella va a _________ pizza para cenar esta noche.</td>
<td>pedir</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>Él puede _________ negocios de manera muy profesional.</td>
<td>dirigir</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>El director quiere _________ unas escenas muy peligrosas.</td>
<td>rodar</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>La presentadora sabe _________ el interés del público.</td>
<td>despertar</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>Los niños deben _________ atención a las palabras del profesor.</td>
<td>poner</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>A mis padres les gusta _________ paseos por las tardes.</td>
<td>dar</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>Si no eres puntual vas a _________ el tren a Madrid.</td>
<td>perder</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>Ella siempre inventa excusas para _________ tiempo y llegar tarde al trabajo.</td>
<td>ganar</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>Cuando él era pequeño, le gustaba _________ bromas a sus amigos.</td>
<td>gastar</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>Él es muy innovador, y pretende _________ camino a otros investigadores.</td>
<td>abrir</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>Cuando sus padres no están en casa, siempre intenta _________ fiestas para ser popular.</td>
<td>montar</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>Los criminales intentan _________ dinero en bancos de Panamá.</td>
<td>blanquear</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>Ella es muy resiliente y consigue _________ fuerzas en los momentos difíciles.</td>
<td>sacar</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>En esta tienda puedes _________ fotos para ponerlas en tu habitación.</td>
<td>revelar</td>
<td>Experimental Incong.</td>
</tr>
<tr>
<td>Para contactar con la compañía, debes _________ el número 948 22 89 04.</td>
<td>marcar</td>
<td>Independent Incong.</td>
</tr>
<tr>
<td>Esta noche voy a _________ las maletas para estar preparado para el viaje de mañana.</td>
<td>hacer</td>
<td>Independent Incong.</td>
</tr>
<tr>
<td>Debes intentar _________ el estrés y no ponerte nervioso.</td>
<td>controlar</td>
<td>Independent Incong.</td>
</tr>
<tr>
<td>Los Youtubers deben _________ vídeos regularmente.</td>
<td>publicar</td>
<td>Independent Incong.</td>
</tr>
<tr>
<td>Todos pueden _________ la web para encontrar información.</td>
<td>navegar</td>
<td>Independent Incong.</td>
</tr>
<tr>
<td>A él no le gusta _________ sus fallos porque es muy arrogante.</td>
<td>reconocer</td>
<td>Independent Incong.</td>
</tr>
<tr>
<td>Los policías tienen que _________ muertes como parte de su trabajo.</td>
<td>presenciar</td>
<td>Independent Incong.</td>
</tr>
</tbody>
</table>
Los estudiantes deben __________ las propuestas antes del miércoles.
La secretaria te ayudará a __________ tus citas con el dentista.
La camarera suele __________ agua a los clientes en el restaurante.
Se pueden __________ cheques en el banco, pero también en el ATM.
El sistema de la web solo permite __________ documentos en formato .pdf.
Los militares necesitan __________ esos puentes para tomar el control.
Las dos compañías consiguieron __________ cuentas tras la negociación.
Su madre le pidió __________ el caos de su habitación.
Los atletas no pueden __________ más millas porque están cansados.
Para __________ una pistola debes tener una licencia de armas.
El padre le intenta __________ la ropa al niño pero él no quiere.
Había mucha gente en el concierto y tenía que __________ la cabeza para ver.
Los amigos decidieron __________ el costo de la comida entre todos.
La forma más segura de viajar es __________ aviones, porque los accidentes son raros.
El novio romántico le dijo antes de viajar que iba a __________ sus besos.
En Europa está permitido __________ bebidas alcohólicas a los 18 años.
El jugador penalizado no podrá __________ partidos hasta noviembre.
El jefe piensa que __________ equipos de trabajadores es muy fácil.
En baseball se deben __________ pelotas con mucha fuerza.
El arquitecto utiliza un software especial para __________ líneas exactas.
La alimentación puede contribuir a __________ diabetes en los niños.

Entregar Independent Incong.
Fijar Independent Incong.
Echar Independent Incong.
Cambiar Independent Incong.
Subir Independent Incong.
Volar Independent Incong.
Ajustar Independent Incong.
Ordenar Congruent
Correr Congruent
Disparar Congruent
Lavar Congruent
Levantar Congruent
Pagar Congruent
Tomar Congruent
Extrañar Congruent
Comprar Congruent
Jugar Congruent
Liderar Congruent
Tirar Congruent
Dibujar Congruent
Desarrollar Congruent
VITA

Manuel F. Pulido Azpíroz

Education

Dual Ph.D., Spanish Linguistics and Language Science 2020
Penn State University,
Department of Spanish, Italian and Portuguese

M.A., Spanish Linguistics 2016
Penn State University,
Department of Spanish, Italian and Portuguese

M.A., Foreign Language Teaching 2008
Universidad de Navarra,
Departamento de Lingüística y Lenguas Modernas

B.A., Spanish Linguistics and Literature 2007
Universidad de Navarra,
Departamento de Lingüística y Lenguas Modernas

Selected peer-reviewed publications


Selected grants and awards

US National Science Foundation Dissertation Improvement Award Fall 2018
($13,610; with Paola E. Dussias)

Language Learning Journal Dissertation Grant ($2,000) Summer 2018

Penn State RGSO Dissertation Support Award ($2,000) Spring 2018