The Metalinguistic Buffer Effect: When Language Comprehension is Good but Linguistic Judgment is Only "Good Enough"

Levi Riven

A Thesis
In the Department
of
Psychology

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Psychology) at
Concordia University
Montréal, Québec, Canada

February 2017

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Levi Riven

Entitled: The Metalinguistic Buffer Effect: When Language Comprehension is Good but Linguistic Judgment is Only "Good Enough"

and submitted in partial fulfilment of the requirements for the degree of

Ph.D., Psychology

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________ Chair
Dr. Joanne Turnbull

______________________________ External Examiner
Dr. Laura Sabourin

______________________________ External to Program
Dr. Charles Reiss

______________________________ Examiner
Dr. Natalie Phillips

______________________________ Examiner
Dr. Karen Li

______________________________ Supervisor
Dr. Roberto G. de Almeida

Approved by: ____
Dr. Karen Li, Graduate Program Director

________________, 2017 ____

Dr. André Roy, Dean of Faculty of Arts and Sciences

ABSTRACT

The Metalinguistic Buffer Effect: When Language Comprehension is Good but Linguistic Judgment is Only "Good Enough"

Levi Riven, Ph.D.

Concordia University, 2017

Native speakers occasionally report incorrect interpretations of language inputs. For instance, when presented with implausible passive sentences, such as *The dog was bitten by the man,* participants sometimes report incorrectly that *the dog* is the agent of the sentence (Ferreira, 2003). According to the good-enough approach, this error occurs because the language comprehension system sometimes relies on rapid heuristics in favour of structure to determine the meaning of a sentence. This thesis tested the hypothesis that the *passive misinterpretation effect* is not an error of comprehension, but arises only *after* interpretation, when participants formulate a thematic judgment of the sentence. It was hypothesized that such thematic judgments involve metalinguistic thinking that can be decoupled from language comprehension processes. This thesis reports the results of two research manuscripts designed to test this hypothesis. In both manuscripts, the methodology involved comparisons of L1 and L2 speakers in single and dual task conditions, which required participants to either report the agents of aurally presented sentences (single task), or to do so while also maintaining six digits in WM (dual task). Because L2 comprehension is more dependent on metalinguistic knowledge than L1 comprehension (Paradis, 2004), it was hypothesized that language status would modulate thematic judgment errors. In Manuscript 1 (Chapter 2), I report the results of two between-subject experiments—one single task and one dual task—which showed that indeed language status modulated the effect. In the single task experiment, there was no difference between L1 and L2 groups. However, in a dual task experiment, the L2 group performed better on the language component of the task (thematic assignment) and worse on the WM component of the task (digit recall). This illustrated that attention allocation to language inputs (at the expense of other tasks) may buttress thematic judgment. In Manuscript 2 (Chapter 4), I report the results of a within-subjects experiment with bilinguals designed to stabilize attention allocation to the different task components, and better isolate the language processing mechanisms that differentiate L1-versus-L2 performance. The data provided evidence for a dissociation between semantic composition and metalinguistic processing in the form of a language-by-load crossover interaction. Bilinguals were better at retrieving the agents of passive sentences in their L2 in the single task condition, but were worse in their L2 in the dual task condition. The pattern suggested that thematic judgments are supported by metalinguistic processes. L2 performance is better in the single task condition because L2 comprehension entails higher initial engagement of metalinguistic representations. In contrast, L1 comprehension depends on implicit semantic compositional processing, which engenders switch costs associated with initiating metalinguistic judgments after interpretation. Critically, because metalinguistic processing is dependent on the control system, the WM load in the dual task condition interfered selectively with L2 comprehension. Thus, L2 performance declined precipitously, while L1 processing remained stable. I conclude that in a native language, “good-enough” heuristics bias metalinguistic conceptualizations of thematic roles post interpretation, and that the underlying semantic composition of passive sentences is fully and faithfully achieved upon an initial analysis by the linguistic system.

Acknowledgements

I would like to thank Dr. Roberto de Almeida, whose mentorship has been instrumental to my development as a scientist, and especially to the work that comprises this thesis. Thank you for the opportunity to conduct this research, for providing all of the necessary resources, for inspiring continued interest in cognitive science and psycholinguistics, and most importantly, for your knowledge, ideas, and constructive feedback, which have guided every stage of this project. I would also like to thank Dr. Norman Segalowitz for his guidance and collaboration in the preparation of Manuscript 1 (Chapter 2) of this thesis. Thanks also to Dr. Natalie Phillips and Dr. Karen Li for providing feedback during the planning and writing stages, and for offering their time and support as members of the internal examination committee.

A special thanks goes to my colleagues at the Psycholinguistics and Cogitation Lab. I am particularly grateful to the many volunteers who helped run this project: Caitlyn Antal, Kimberly Bless, Kaegan Brisebois Kancachian, Maude Brisson-McKenna, Alexa Ray Falcone, Geraldine Jeckeln, Lisa Martignetti, Gabrielle Mourra, Alexandra Therond, Vanessa Vassiloganis, and Katharina Wegner. Thank you all for your help and for being such reliable and productive research assistants.

To my fellow graduate students and friends—John Brand, Hilary Duncan, Elizabeth Hebert, Kristyn Wright, Danit Nitka, and Jessica Senn—who punctuated these years of study with repartee, levity, and the occasional spit-take, thanks for making it fun. And to my siblings and siblings-in-law, thank you for all the support and comfort food, and for allowing me to practice psychology on your children. (They are all fascinating and adorable subjects.) Thanks also to *RC* for your encouragement and the many interesting chats about language. Most of all, to my father, Pinchus Riven, thank you for your love, wisdom, and unwavering support; for your ongoing enthusiasm about my work, and for keeping me on my linguistic toes throughout my life with your unrelenting puns. This accomplishment feels ever more meaningful knowing that it makes you proud.

Dedications

To the memory of my sister *Ada Riven-Herzog* (1976—2015), whose smile withstood the many hardships of her life, and whose knack for deep analysis made her so much fun to talk to.

~

To the memory of my mother *Beryl Lapin-Riven* (1950—2004), who initially sparked my curiosity for language with her affectionate renderings of children's literature, and whose enthusiasm for sounds, words, and phrases continues to fuel that curiosity. This thesis opens with a riddle she often told. So once again, she gets the credit for nurturing my first words.

Contribution of Authors

Chapter 2 of this dissertation presents a research manuscript that was co-authored by Dr. Roberto de Almeida and Dr. Norman Segalowitz. For this manuscript, I formulated the research question, conceptualized the design, coordinated data collection, performed the statistical analyses, and wrote the initial draft. All of this work was conducted under the supervision and guidance of Dr. Roberto de Almeida and in consultation with Dr. Norman Segalowitz, who both contributed to the experimental design and provided critical revisions to the manuscript.

Chapter 4 of this dissertation presents a research manuscript that was co-authored by Dr. Roberto de Almeida. As in the previous manuscript, I conceptualized the research question and experimental design, coordinated data collection, performed the statistical analysis, and wrote the initial draft. Dr. de Almeida supervised all stages of this work, contributed to design conceptualization and provided critical revisions to the manuscript.

Table of Contents

List of Figures

List of Tables

Chapter 1: General Introduction

The Metalinguistic Buffer Effect: When Language Comprehension is Good but Linguistic Judgment is only "Good-Enough."

Consider the following riddle: 'A man looks at a picture and says, *"This man's father is my father's son. Brothers and sisters I have none."* Who is the man in the picture?' Although it is impossible to determine the answer to this question without fully processing the structure and meaning of the linguistic input, sentence processing alone is insufficient for finding the solution. Problem solvers must not only parse the syntax of the riddle's constituent clauses; they must also *reason* about what is entailed by the interpretations they derive. It is perhaps unsurprising that processing usually ensues with less than perfect success. While the answer to the riddle is the narrator's son, individuals often surmise that it is the narrator instead.

Now consider another kind of question: *Who is the agent of the sentence* "*The dog was bitten by the man"?* Although this question is certainly easier than the first, the two puzzles share a number of relevant characteristics. For both, the solution is entailed by the structure and meaning of the linguistic input; for both, the input structures are grammatical; and for both, processing ensues with less than perfect accuracy. Whereas *the man* is the agent of the sentence in question, native speakers of English sometimes report that it is *the dog* instead (Ferreira, 2003). But while it is obvious that failing to solve the first puzzle reflects breakdowns in reasoning rather than language processing, it is much less clear in the latter case whether it is *interpretation* per se, or some form of post-interpretive *thinking* process that goes awry.

The objective of this thesis is to present a methodological approach—involving comparisons of first and second-language performance—for dissociating these two kinds of information processes within the context of a similar thematic judgment task. Given the evidence collected from two studies, I will propose that the agentive misinterpretation effect reported by Ferreira (2003) reflects metacognitive interference with linguistic outputs, and does not result from incomplete comprehension processing, as the good-enough language framework suggests (Karimi & Ferreira, 2015).

The Good-Enough Approach to Language Processing

According to the good-enough approach (Ferreira, Bailey, & Ferraro, 2002), sentence processing routinely falls short of producing a fully specified and faithful representation of the input. For instance, when asked "*How many animals of each kind did Moses bring onto the ark?"*, native speakers who are familiar with the biblical narrative frequently answer "two" without noticing the false supposition behind the question: it was Noah, not Moses, who populated the ark (Erickson & Mattson, 1981). Failures to detect such anomalies suggest that interpreters do not always retrieve the full meaning of each word in a sentence and may ultimately build representations that betray the semantics of the input (see Barton & Sandford, 1993 for a similar demonstration). Ferreira and colleagues (Ferreira et al., 2002, Ferreira & Patson, 2007) refer to such errors as "good-enough" representations to emphasize their adaptive nature. According to the approach, the language comprehension system is computationally miserly, and therefore deploys "fast-and-frugal" heuristics to achieve minimally sufficient interpretations. While these heuristics may be prone to errors with certain kinds of inputs, they are thought to be correct most of the time, and useful for sparing the comprehension system from completing effortful compositional analyses (Ferreira & Patson, 2007).

Christianson, Hollingworth, Halliwell, and Ferreira (2001) have demonstrated that good-enough processing may occur within the domain of syntax as well, specifically, for the interpretation of garden-path sentences. Garden-paths, such as *While Anna dressed the baby played in the crib,* are difficult to parse due to ambiguity concerning the attachment of the internal noun phrase (NP) *the baby.* While the NP initially appears to be the object of the opening adverbial clause *While Anna dressed,* it can only be assigned as the subject of the main clause *played in the crib* when the sentence is considered as a whole. Thus, readers who initially assign the NP to the first clause must later revise the attachment when the second clause is encountered. To the extent that revision is successful, the language comprehension system yields the following interpretation: *While Anna dressed [herself], the baby played in the crib.* Although garden-paths are widely recognized as being computationally demanding—owing to the need for revision—some approaches to sentence processing assume that the parser eventually recovers from the initial attachment and ultimately outputs a fully specified structural rendering of the phrase (e.g., Fodor & Inoue, 2000). However, Christianson et al. (2001) showed that when participants are asked "*Did Anna dress the baby?"* and "*Did the baby play in the crib?",* they responded affirmatively to both questions, suggesting that they had never fully resolved the ambiguity. Instead, they appeared to cling to both the initial attachment error and the subsequent reparation. Thus, just as in the case of semantic retrieval, structural processing appears to be *good enough* to output a workable interpretation, but perhaps not sufficiently *good* to guarantee a fully specified rendering of structure all of the time.

Recently, Frazier and Clifton (2015) demonstrated yet another good-enough processing illusion for syntactically blended sentences such as *A passerby rescued a child from almost being run over by a bus.* Such blends resemble speech disfluencies that result from interference between two independently planned utterances—i.e., "*A passerby rescued a child from being run over…*" and "*A child was almost run over by a bus*". Speech disfluencies like these serve as a central motivation for the good-enough approach (Ferreira & Bailey, 2004). Given that disfluencies are common—occurring 6-10 times for every 100 words in natural speech contexts (Bortfeld, Leon, Bloom, Schober, Brennan, 2001)—Ferreira and Bailey (2004) suggested that the language comprehension system likely adapted principled strategies for repairing disfluencies. With respect to syntactic blends, the comprehension system may restore the input to the probable intentions of the speaker by outputting a coherent interpretation that overwrites the disfluent component; and this may result in failures to notice the strange entailments of the sentence—i.e., that the child must have been run over as a consequence of having been rescued from an *almost* tragic encounter with the bus. Indeed, Frazier and Clifton (2015) found that participants fail to notice the disfluency: they judge these kinds of blended utterances as acceptable—on par with unblended controls—and they do so with near ceiling frequency (m=.97).

The picture that emerges from these kinds of demonstrations, among others (Sandford & Sturt, 2002; Wason & Reich, 1979), is that interpreters may not always achieve a fully specified word-for-word interpretation in accord with the compositional meaning of sentences. In some instances, the language comprehension system appears to rapidly output an interpretation that is more consistent with what the speaker most likely intended to say, rather than with what was in fact uttered (Barton & Sandford, 1993; Erikson & Mattson, 1981; Frazier & Clifton, 2015; Wason & Reich, 1979). In other instances, the language comprehension system appears to forgo costly reanalyses required to disambiguate structures, and instead may output partially parsed fragments that are incompatible with the full scope of the input (Christianson et al., 2001; see also Sanford & Sturt, 2002).

The good-enough approach offers a dual-process account of language processing that aims to unify these diverse phenomena (Karimi & Ferreira, 2015). Ferreira and colleagues propose that the human parsing mechanism is comprised of two independent processing streams—one heuristic and one algorithmic. The heuristic stream deploys "fast and frugal" strategies for building highly probable interpretations of sentences that are sometimes unfaithful to the input structure. The algorithmic stream, in contrast, executes effortful syntactic computations that yield fully compositional representations (see Townsend & Bever, 2001 for a similar proposal). According to the good-enough approach, interpretation errors occur because the parser operates according to the "*minimal effort* principle" (Ferreira & Patson, 2007, p.80): that is, the comprehension system aims to select the most rapidly available, sufficiently satisfactory interpretation, and therefore, may truncate the computations of the syntactic-algorithmic stream as soon as a "good-enough" representation is available (Karimi & Ferreira, 2015).

The Case of (Implausible) Passives

Of all the phenomena reviewed by Ferreira and colleagues under the rubric of good-enough processing (Ferreira et al., 2002; Ferreira & Patson, 2007; Karimi & Ferreira, 2015), one kind of error stands out from the rest because it involves apparent misinterpretations of unambiguous, grammatical, and relatively common structures—i.e., simple transitive passives, such as (1a) and (1b). In Ferreira's (2003) paradigm, participants listened to such plausible and implausible passives, as well as their active controls (see 1c-d), and were cued at sentence offset to report the agent (and patient) noun of these sentences. Ferreira (2003) found that participants reported the agentive noun correctly most of the time (all means ≥ 74%); however, whereas they performed at ceiling for active sentences (regardless of plausibility), they misconstrued the agents of plausible passives such as (1a) 12% of the time, and of implausible passives such as (1b) 26% of the time.

(1) a. The man was bitten by the dog. *Plausible passive*

b. The dog was bitten by the man. *Implausible passive*

c. The dog bit the man. *Plausible active*

d. The man bit the dog. *Plausible passive*

Ferreira (2003) reasoned that because interpretations were correct in the majority of instances for all sentence conditions, participants surely had the requisite grammatical knowledge to correctly interpret actives and passives alike. However, the fact that they erred some of the time suggested moreover that they also used non-compositional processing strategies, which interfered with the structurally licensed interpretation. Bever (1970) was perhaps the first to describe the heuristics that Ferreira (2003) and colleagues (Christianson, Luke, & Ferreira, 2010) later adopted in their good-enough model. The most relevant heuristic concerns the linear ordering of sentence constituents. In English, agent and patient roles typically map onto the first and second main NPs of a sentence, providing a superficial cue for determining who did what to whom. As Bever (1970, p.299) put it, “any [noun-verb-noun] sequence in the surface structure is assumed to correspond to *actor-action-object* [or agent-verb-patient] in the underlying structure.” Thus, to the extent that interpreters rely on the unmarked order of constituents—i.e., *dog bite man*—to determine how the arguments of a verb are interrelated, interpretations will go awry for passives, which bear a non-canonical *patient-verb-agent* sequence.

Another heuristic that is relevant to Ferreira’s (2003) results is what Bever (1970, p.296) referred to as the “semantic strategy”, that is, the tendency “to combine lexical items in the most plausible way… independent of syntactic structure.” The semantic plausibility heuristic occurs when interpreters’ *a priori* intuitions about how the world works influence how they interpret sentences. Thus, as Ferreira showed, the tendency to misconstrue the thematic roles of passive sentences was exacerbated when the events being described were implausible, presumably because participants had accessed a plausible event representation contradicting the input structure.

Consistent with the good-enough approach, the passive misinterpretation effect provides evidence for the presence of at least two different kinds of cognitive strategies involved in sentence processing—one superficial and one fully compositional. But the precise locus and mechanism of the error—that is, whether it occurs as a direct function of semantic compositional processes—is not entirely clear. While it seems evident that linguistic inputs activate heuristics that bias participants’ off-line judgments, this does not require that the parser truncates syntactic analysis prematurely, or that structural processing more generally is effortful, and slower than other aspects of semantic composition. In the subsequent section, I will present an alternative mechanism for the effect within the context of a broader literature review on passive sentence processing.

What is Difficult about Processing Passive Sentences?

The notion that passive sentences are structurally complex and difficult to process can be traced back to the dawn of the cognitive revolution, with Chomsky's (1957) landmark publication *Syntactic Structures*. In *Syntactic Structures* Chomsky proposed that the diverse set of grammatical (English) sentences are all derived from a finite set of axioms that apply to more simplex structures, called *kernels.* For instance, the sentences in (2b-d) can all be derived by applying a set of relevant transformations to (2a). Thus, the *passive transformation* is applied to strings like (2c) and specifies how to derive the passive sentence from the kernel (2a)—that is, by "interchanging the two noun phrases, adding *by* before the final noun phrase, and adding *be+en* to *aux*" (Chomsky, 1957, p.61).

(2) a. John bites Mary. *Kernel: Affirmative Active*
b. John does not bite Mary. *Negative*
c. Mary was bitten by John. *Passive*
d. Mary was not bitten by John. *Negative Passive*

In the 1960s, psychologists interested in adopting transformational grammar as a model for syntactic processing proposed that the perceptual complexity of a sentence is a function of the number of transformations required for its derivation (McMahon, 1963; Mehler, 1963; Miller, 1962, Miller & McKean, 1964; Slobin, 1966). Several early experiments appeared to support the theory, by showing for instance that participants take longer to transform the *negative passive* in (2d) from the *active* in (2a) than they do to transform the sentence from more closely related structures, such as (2b) or (2c). The passive transformation in particular was estimated to take about 1.2 seconds (Miller, 1962; McKean & Miller, 1966).

Different kinds of methods were used to illustrate the derivational principle with respect to passives. For instance, Miller and McKean's (1964) results reported above were based on the speed with which participants matched passives in one list of sentences with their corresponding actives in another list of sentences (compared to other hypothesized transformations). Slobin (1966) used images that depicted simple actions and had participants verify whether active and passive descriptions matched the action. (He found that passive descriptions took longer to verify than active descriptions.) Mehler (1963) used a recall task involving active and passive sentences (and others) and revealed that sentences tended to be recalled in the active voice.

Although these kinds of data were initially taken to support the derivational theory of complexity (DTC), later observations illustrated problems with the idea that syntactic processing difficulty was a simple function of transformational distance (Bever, 1970; Fodor & Garret, 1967). Fodor and Garret (1967), for instance, pointed out that (3a) is easier to process than (3b) even though the former is a derivation of the latter. Fodor and Garrett suggested instead that the processing difficulty of a sentence depends more directly on the nature of its superficial perceptual properties, which provide clues for how the input string maps onto its structure. Thus they showed that although (4a) and (4b) have a common structure, and although (4a) is shorter than (4b), the former is more difficult to process because the deletion of the relative pronoun *whom* removes a useful perceptual cue for recognizing the sentence's embedded clause.

(3) a. The boy was hit.
b. The boy was hit by something.

(4) a. The man the dog bit died.
b. The man whom the dog bit died.

One of the issues with early studies of DCT was that they relied heavily on off-line measures—such as long-term recall (Mehler, 1963) and search-and-match speed (Miller & McKean, 1964)—which engage different kinds of cognitive processes extraneous to parsing and interpretation. In the ensuing years, a variety of on-line processing studies showed that passives were in fact just as easy—and by some measures easier—to process than actives (Carrithers & Bever, 1984; Carrithers, 1989; Frazier, Taft, Roeper, Clifton, & Elrich, 1984). Specifically, the self-paced processing rates per letter (Carithers & Bever, 1984) and per word (Carrithers, 1989; Frazier et al., 1984) were faster for passives than for actives, and these kinds of effects could not be accounted for by the greater density of function words in passives (Carrithers, 1989).

It is beyond the scope of this review to explore all the possible causes for the discrepancy between off-line and on-line measures of passive processing. But one possibility that is particularly relevant for our purposes is that passive sentences may have systematic effects on how people think about the information they contain, independent of what is required to process their structure. Consider that the passive voice can be used to modulate how attention is drawn to specific aspects of an event. As the sentences in (5a-c) illustrate, the passive voice allows communicators to stage events from the perspective of their objects (or patients) and omit their subjects (or agents) entirely. The result is that passives place the patient noun directly in the attentional spotlight and steer the agentive noun into the periphery—a handy device for communicators who wish to reduce the salience of agents. Hence, numerous politicians—including Richard Nixon, Henry Kissinger, Bill Clinton, Chris Christie, and at least three Bushes (Fallows, 2015)—have relied on some variation of (5b) to shield themselves from blame when making apologetic statements (Pinker, 2014).

(5) a. The boy was hit.

b. Mistakes were made.

c. Participants were instructed to listen carefully to each sentence before providing a response.

Pinker (2014) noted that although the passive voice is frequently disparaged by prescriptive grammarians, it is in fact more appropriate stylistically and indeed more felicitous than actives under certain circumstances. Consider the following illustration provided by Clark and Clark (1968, crediting Morton, 1966): the reader will likely find the two-sentence sequences in (6a) and (6b) to be quite natural, but those in (7a) and (7b) to be somewhat odd. One apparent determinant of felicity in these examples is whether the second sentence in the sequence places the topic of the discourse at the first or second noun position. Because English speakers place what they are talking about at the beginning of a sentence, readers prefer when the topic is positioned there, regardless of its thematic status. Thus, (6a) seems more felicitous than (7a) even though it uses the passive voice and mentions the patient noun first.

(6) a. I saw the house. The house was built by the man.

b. I saw the man. The man built the house.

(7) a. I saw the house. The man built the house.

b. I saw the man. The house was built by the man.

These kinds of examples illustrate that usage norms can skew how we think about the events that sentences encode, independent of how we process their syntax or thematic roles. Returning to Ferreira's (2003) passive misinterpretation effect, it seems plausible that participants may err not because they have difficulty parsing passive structures or because they misinterpreted the sentence, but rather because opening nouns provide the focal point for how they think about the event the sentence describes. Pragmatically, a sentence such as *The dog was bitten by the man* places *the dog* within the focus of attention. Thus, when participants are asked to report the agent of this sentence, they are required to inhibit a front-and-center representation in order to retrieve a less prominent constituent. Even if their interpretation is correct, they may experience difficulty generating the correct response.

This difficulty is compounded by the fact that the response criterion—i.e., the agentive noun—is a metalinguistic construct, and provides only an indirect index of interpretation. In the natural course of language processing, interpreters do not consciously track elements of the input in terms of their grammatical, semantic, or thematic classes. Rather, they deploy implicit processing strategies to rapidly build up a conceptual representation. Thus in order to report the agent of a sentence, the interpreter must initiate a post hoc metalinguistic formulation of the output. While these formulations are certainly informed by sentence interpretations, the response generation process per se occurs outside the realm of sentence processing. According to the good-enough approach, interpretations go awry because linear-order and plausibility heuristics interfere with fully compositional comprehension processes. But this claim assumes that the off-line response criterion is the direct output of the language comprehension system and overlooks the intervening metalinguistic processes that must be initiated after interpretation to formulate a response. The alternative hypothesis that will be tested here is that heuristics interfere with post-hoc metalinguistic formulations of the output, not parsing or interpretation. As detailed in the preceding paragraphs, attentional focus may serve as the primary mechanism by which linearity and plausibility cues derail metalinguistic judgments.

The Current Program of Research

The overarching aim of this program of research is to provide evidence that (a) off-line thematic assignment entails metalinguistic thinking, which can be dissociated from parsing and interpretation, and (b) linearity and plausibility heuristics interface with the metalinguistic stage of processing. The methodology used here takes advantage of differences between first and second language (henceforth, L1 and L2) processing. The rationale is as follows: L1 and L2 processing is marked by a fluency gap that affects the automaticity of language comprehension processes (Segalowitz, 2010). Because first and second languages differ in terms of age of acquisition, quantity and quality of exposure, circumstances of learning, and regularity of usage, differences emerge in how the underlying language representations are stored and used (Paradis, 2004; 2009; Ullman, 2001). For the present purposes, the most relevant difference concerns the degree to which parsing and interpretation relies on implicit semantic composition strategies (L1>L2) or explicit metalinguistic processing (L2>L1).

Metalinguistic processing refers to any cognitive analysis that draws on what the interpreter consciously knows about the norms and rules of language use. While both L1 and L2 speakers may use metalinguistic knowledge to some extent depending on the circumstances, L2 speakers are more dependent on metalinguistic processing for comprehension. This is partly because L2 acquisition entails explicit memorization of grammatical rules, and partly because most L2 speakers never fully automatize the implicit semantic composition strategies that drive L1 parsing and interpretation (Paradis, 2009). These L1-L2 differences affect how the control system is engaged in language interpretation (Paradis, 2009). Implicit semantic composition processes allow native speakers to automatically integrate structural information on-line (Clahsen & Felser, 2006), but may also entail delays or costs in activating the control system when metalinguistic thinking is required. In contrast, the relative absence of implicit processing strategies in L2 necessitates higher initial engagement of the control system during L2 interpretation, potentially easing access to metalinguistic representations when deliberation is permitted (Paradis, 2009), but impeding interpretation when it is not (Clahsen & Felser, 2006).

The general method of this research involved comparisons of L1 and L2 thematic assignment in single and dual task conditions. The single task paradigm was similar to Ferreira's (2003) task—i.e., participants listened to active and passive sentences, that were either plausible or implausible, and were cued at sentence offset to report the agentive noun. The dual-task paradigm involved the same task but with the addition of a six-digit working-memory (WM) load. It was hypothesized that single-task performance would provide evidence for a *metalinguistic buffer effect*—specifically, L2 speakers would perform better than L1 speakers due to higher initial activations of metalinguistic knowledge during interpretation. However, because metalinguistic processing is dependent on the control system, it was further hypothesized that patterns of L1-L2 performance would be reversed in the dual task condition. That is, L1 interpretations would persevere to a greater degree under load given the availability of automatic semantic composition processes. To the extent that such predictions are obtained, the data would provide evidence that (a) L1 compositional parsing and interpretation is automatic, rather than effortful and slow, (b) off-line thematic assignment is an index of both linguistic and metalinguistic processes, and (c) thematic assignment errors may occur primarily as a function of metalinguistic processing dynamics. That is, what is effortful about reporting the agents of passive sentences is that it requires metalinguistic formulations, which may be skewed by how linearity and plausibility cues draw attention to event constituents.

As a brief aside, although it may seem unlikely that L2 speakers would outperform their L1 counterparts on language processing tasks, several studies have reported such effects with respect to off-line thematic assignment (Chipere, 1998; Debrawska & Street, 2006; Sasaki, 1998). Although such results were not anticipated, these researchers all attributed their findings to L2 metalinguistic processing. To my knowledge, the experiments reported here are the first to investigate the *metalinguistic buffer effect* prospectively, and to test specific hypotheses concerning its underlying mechanisms.

Preview of Studies

The current program of research evolved in two stages, which were reported in two research manuscripts. Manuscript 1 appears in Chapter 2 and Manuscript 2 appears in Chapter 4. (Chapter 3 bridges these two experimental reports.) Both manuscripts report research that aimed to test the above hypotheses, but with slight differences in methodology. Specifically, Manuscript 1 (Chapter 2) reports two between-subject experiments comparing L1 and L2 speakers agentive assignments in either a single task (Experiment 1) or a dual task (Experiment 2) paradigm. The results of these experiments showed that off-line thematic assignment draws on generalized controlled processes, which in turn, established that off-line agentive reports are not a direct index of semantic compositional processes. Specifically, we found a task selection effect in the dual-task experiment, whereby L2 speakers performed better than L1 speakers on thematic assignment, but worse on digit recall. However, owing to group differences in task selection, the specific mechanisms underlying L1-L2 differences could not be adequately isolated.

To address this limitation, the second stage of research (reported in Manuscript 2, Chapter 4) employed a within subjects-design with bilinguals. Here, the L1-L2 comparison was manipulated in terms of language of presentation, rather than language group, to reduce the likelihood of between-language task selection differences. In addition, we integrated additional measures into the design, such as reaction times and WM capacity, to better isolate the underlying cognitive mechanisms of L1-L2 differences. This experiment supported the hypotheses in the form of a task-by-language crossover interaction. As predicted, bilinguals were more successful assigning agents in their L2 in the single task condition, but were more successful in their L1 in the dual task condition. Additional L1-L2 effects associated with reaction times and WM capacity in the single-task condition further supported the specific mechanisms hypothesized above. In Chapter 2, Chapter 4, and the General Discussion (Chapter 5), I outline the implications of these results for the good-enough model and for our understanding of sentence processing more generally.

Chapter 2: The Role of the Centralized Control System in the "Good-Enough" Processing of Passives

Sentence comprehension has been said to rely heavily on the correct assignment of thematic roles to the arguments of a verb (Frazier, 1987; Tanenhaus, Carlson, & Trueswell, 1989; see de Almeida & Manouilidou, 2015, for a review). For instance, given the sentence *The boy pushed the girl,* the reader or listener must determine that the boy is the *agent* of the sentence—the one doing the pushing—and that the girl is the *patient*—the one being pushed. Although interpreting these roles correctly is presumably necessary for understanding a sentence, Ferreira (2003) showed that native speakers frequently misconstrue the roles associated with a verb's subject and object nouns. When presented with implausible passive sentences such as *The dog was bitten by the man*, participants sometimes reported incorrectly that *the dog* was the agent of the sentence—the *biter*, not the patient *bitee* (Christianson, Luke, & Ferreira, 2010; Ferreira, 2003).

This result is particularly surprising because the sentences Ferreira and colleagues employed were grammatical and unambiguous. Therefore, the question arises as to why a native speaker who is said to fully process the structure and meaning of a sentence—often automatically or, as Fodor (1983) put it, like a reflex—should derive incorrect interpretations. In their good-enough approach to language processing, Ferreira (2003) and colleagues (Christianson et al., 2010; Karimi & Ferreira, 2015) offered a different characterization of the language comprehensions system—one in which structure and meaning is often underspecified due to the difficulty associated with fully processing all the elements of the input. The model suggests that rather than computing a detailed syntactic parse of a phrase, the comprehension system sometimes relies on extra-grammatical heuristics to achieve a "fast and frugal" but possibly erroneous interpretation.

In support of the good-enough approach, a variety of experiments have shown that participants sometimes report interpretations of utterances that deviate from the syntactically licensed meaning (e.g., Christianson et al., 2010; Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira, 2003; Swets, Desmet, Clifton, & Ferreira, 2008). In general, these errors occur due to domain-specific sources of processing difficulty—including structural complexity (Ferreira, 2003), structural ambiguity (Christianson et al., 2001), and syntactic blends that create logical contradictions (e.g, *The boy was rescued from almost drowning;* Frazier & Clifton, 2016). The picture that emerges from these findings is that the dynamics of veridical versus good-enough interpretations are largely a function of challenges inherent in the signal. An open question for the good-enough approach concerns the extent to which misinterpretations vary with domain-general, or non-linguistic, sources of processing difficulty. Here we investigate the passive misinterpretation effect in native and non-native speakers in single- and dual-task experiments. Our aim is to further operationalize the factors that influence good-enough misinterpretations in terms of domain-general constraints on processing.

The Good-Enough Approach to Thematic Assignment

At least two heuristics are thought to skew the rendering of thematic roles in non-canonical sentences, one computed from linearity cues and the other from semantic associations (Bever, 1970; Caramaza & Zurif, 1976; Ferreira, 2003; Townsend & Bever, 2001). To illustrate, consider the findings in Table 1. Each of these sentences describes a biting incident involving a dog and a man, but vary with respect to syntactic structure (active versus passive voice) and semantic plausibility (dog-as-biter versus man-as-biter). In English, agent and patient roles typically map onto subject and object positions respectively. Thus, one might rapidly construe these roles by simply attending to the unmarked order of constituents—dog-bite-man—without fully processing the structure. Indeed, when asked to report the agent of sentences like those in Table 1, participants performed near ceiling as long as the agent occupied the canonical position (a) and (b). But when thematic roles were inverted in passive sentences (c) and (d), errors increased, with greater confusion between the roles assigned to the verb's arguments. In the case of implausible passives (d), in particular, about one fourth of the participants report that the patient (dog) was the agent of the event.

The semantic content of the arguments and the plausibility associated with their relations may also provide cues for determining the agent of a sentence. Given a man, a dog, and a biting incident, the dog is typically the perpetrator, and this knowledge biases the comprehension system towards a more plausible interpretation. Hence, in Ferreira's (2003) and Christianson et al.'s (2010) experiments, the tendency to invert thematic roles was exacerbated when sentences described an implausible event. Collectively, the results in Table 1—as well as those from earlier experiments (e.g., Bever, 1970; Caramazza & Zurif, 1976; Fodor, Bever, & Garrett, 1974)—

Table 1

Proportion correct assignment of thematic roles from Ferreira (2003) and Christianson et al. (2010).

Condition	Example Sentence	Ferreira (2003)	Christianson et al. (2010)
a. Active Plausible	The dog bit the man.	0.99	0.99
b. Active Implausible	The man bit the dog.	0.99	0.91
c. Passive Plausible	The man was bitten by the dog.	0.88	0.90
d. Passive Implausible	The dog was bitten by the man.	0.74	0.74

indicate that native speakers often deploy heuristics in thematic assignment tasks resulting in erroneous interpretations.

Ferreira (2003) and Christianson et al. (2010) note that thematic assignment is not a simple function of non-syntactic cues. Central to the good-enough approach is the notion that the structural parameters of the input are a critical source of interpretation. Table 1 shows that even for implausible passive sentences, participants retrieved the correct noun most of the time, suggesting that a structural analysis of the input usually prevails.

Ferreira (2003) thus postulated independent syntactic and heuristic processing streams to account for the full pattern of effects (see Townsend & Bever, 2001, for a similar proposal). The heuristic stream draws on linear order and plausibility cues to compute a rapid interpretation, whereas the syntactic route applies more taxing structural algorithms. Most often, the outputs of the two streams concur and interpretation proceeds felicitously. But when heuristic and algorithmic interpretations conflict, costly analyses are required to integrate the diverse elements of the input (Lim & Christianson, 2013a). According to the good-enough framework, errors ensue because the comprehension system will frequently forgo effortful syntactic analyses if a "good-enough" interpretation is available in the early stages of processing (see Ferreira & Patson, 2007, and Kamiri & Ferreira, 2015, for reviews).

What is the Nature of 'Veridical' Processing in the Good-Enough Model?

The core thesis of the good-enough approach is that syntax and heuristics provide alternative means of determining how the elements in a phrase are interrelated, and that non-veridical interpretations occur when the heuristics prevail over syntax. But the model is less clear about the relation of the syntactic mechanism to the veridical response. In general, the good-enough approach operationalizes syntax as algorithmic and computationally demanding. In some versions of the architecture, the algorithmic stream is said to be responsible for marshalling the output of faster, and possibly earlier, heuristic processes (Karimi & Ferrera, 2015). Elsewhere, correct interpretation is said to require integration of competing syntactic and heuristic outputs (Christianson et al., 2010; Lim & Christianson, 2013a, b), but the nature of these integration processes—that is, whether they are constitutive of or extraneous to the parsing mechanism—is left open. Thus, it is unclear whether interpretation failures arise from (a) difficulty in parsing the structure initially, including the building of a logical form, (b) difficulty in integrating competing structural and non-structural outputs at a subsequent stage of parsing, and/or (c) difficulty in coordinating a response to an experimental probe after the linguistic input has been altogether dismissed by the parser. It is therefore also unclear whether veridical interpretations arise from (a) the perseverance of an initial syntactic parse over spurious cues for agency, (b) mechanisms that adjudicate between the various cues during parsing, and/or (c) the engagement of non-linguistic control mechanisms after parsing.

One means of clarifying the relevant mechanisms is by operationalizing performance on Ferreira's (2003) thematic assignment task in terms of domain-specific versus domain-general constraints on processing. By domain-specific we refer to processes that are unique to the human parsing mechanism, such as the on-line computation of a phrase structure rule and, consequently, semantic composition that is faithful to the linguistic input. By domain-general we refer to processes that underlie multiple cognitive systems, such as selective attention and executive control functions. Both parsing-specific and generalized control functions are necessary for natural language processing (Fedorenko, 2014; Fedorenko & Thompson-Schill, 2014), but their roles in comprehension may vary with the nature of the processing constraints imposed by a task—specifically, whether task demands load on one system versus the other. Interpretation tends to default to automatic parsing mechanisms when domain-general resources are taxed (e.g., by a working-memory load), and comprehension relies increasingly on domain-general control processes when the complexity of the input, or the means by which the response criterion is measured, overloads the incremental operations available to the on-line parser.

This mutually compensatory nature of domain-specific and domain-general processes was demonstrated in a dual-task experiment (Fedorenko, Gibson, & Rohde, 2006). Fedorenko et al. showed that the speed and accuracy of sentence processing interacts with short-term verbal recall as a function of both increasing syntactic complexity and memory load. Participants performed worse on verbal recall when interpreting object-extracted sentences such as (1a) compared to subject-extracted sentences such as (1b)—of which the former is more syntactically complex due to a non-local dependency in the embedded clause (i.e., the object of "consulted" is not adjacent to the verb). And participants performed worse on sentence processing measures—particularly for the object-extracted clauses—when the memory load increased from one noun to three. The dynamics of these trade-offs suggest firstly that on-line interpretation of simple syntactic forms can proceed relatively independent of the general control system—i.e., processing of subject-extracted clauses interacts less with a non-syntactic processing load.

However, as the complexity of the input structure increases, control processes become more and more critical to comprehension.

(1) a. The physician who the cardiologist consulted checked the files in the office.

b. The physician who consulted the cardiologist checked the files in the office.

Returning to our earlier question, we can better demarcate the dynamics of veridical versus good-enough interpretation by exploring how accuracy varies as a function of domain-general constraints on processing. Ferreira (2003) and Christianson et al. (2010) have already shown that accuracy declines with domain-specific factors, such as structural complexity and the presence of competing cues for agency in the signal. But on the basis of their findings alone it is unclear what occurs once the parser faces these challenges. Does comprehension default to whatever output is initially computed by the parser? Or does it also reflect some kind of centralized effort? For instance, the interpreter may have to inhibit the allure of linearity and semantic cues and/or review syntactic markers for agency off-line in order to respond correctly to the experimental probe. In two experiments reported here, we provide evidence that indeed centralized processes support accurate responding in the thematic assignment task. Specifically, we show that when a peripheral task constrains domain-general resources during processing, interpretation accuracy varies with how native and non-native speakers allocate attention between the linguistic and non-linguistic tasks. In the subsequent section, we briefly discuss good-enough processing in first- (L1) and second-languages (L2) and elaborate on the rationale of our experiments.

Good-Enough Language Processing in L1 and L2

It is frequently noted that L1 and L2 sentence processing differs in a variety of ways (Clahsen & Felser, 2006; Paradis, 2004; 2009; Ullman, 2001; 2015). Although there is fierce debate as to whether differences reflect a single linguistic system operating at different levels of proficiency (Dekydspotter, Schwartz, & Sprouse, 2006), or reflect the deployment of fundamentally different processing systems (Bley-Vroman, 1990; Paradis, 2009; Ullman, 2015), it is beyond controversy that L2 processing is marked by a fluency gap (Segalowitz, 2010), which affects how linguistic data is integrated during comprehension (Christianson, 2016). Because L1 and L2 speakers of a given language have different experiences with that language—including age of acquisition, circumstances of learning, quantity and quality of exposure, and the presence or absence of a previously learned language—differences may emerge in the speed and fidelity of language comprehension processes.

Although in general the L1-L2 fluency gap favours L1 processing, the vastly different language experiences of native and non-native speakers can sometimes create unexpected boons for the latter population (Debrawska & Street, 2006). For instance, Paradis (2009) noted that although L2 speakers have more difficulty analyzing structure on-line, they tend to have more metalinguistic knowledge of the grammar due to having explicitly memorized grammatical rules during acquisition. Similarly, Christianson (2016) noted that having less experience with a language might minimize the impact of linearity cues on L2 comprehension, as linearity effects may be associated with frequency of exposure. With respect to thematic assignment in particular, L2 interpretation of passives may be relatively less biased by the word-order heuristic. Debrawska and Street (2006) showed for instance that non-native speakers are sometimes more accurate than native-speakers when reporting the agentive noun of passive sentences in an off-line task, which may suggest that indeed, non-native speakers are less sensitive to linearity cues.

Observations such as these imply that the nature of competition between syntactic and non-syntactic markers for agency during comprehension is different for L1 and L2 speakers, but how these differences manifest in performance is not immediately obvious. Factors that are presumed to favour L2 comprehension—such as metalinguistic knowledge and insensitivity to probabilistic elements of the input—may be offset by a tendency to underspecify syntax during on-line processing (Clahsen & Felser, 2006). In turn, non-native speakers may be more biased by semantic cues for agency than their native-speaking conspecifics (Christianson, 2016; Clahsen & Felser, 2006). Thus it was shown that Korean L2 speakers of English made more semantically biased errors when translating English passives into Korean—i.e., L2-to-L1 translations—than when translating Korean passives into English—i.e., L1-to-L2 translations (Lim & Christianson, 2013a).

Although the dynamics of veridical versus good-enough processing may differ for L1 and L2 speakers for a multitude of reasons, in general, successful comprehension in an L2 requires relatively more controlled processing effort (Paradis, 2004; 2009; Segalowitz, 2010; Ullman, 2001). Thus, situations in which L2 speakers outperform native speakers are generally restricted to tasks that permit deliberation prior to responding (e.g., Debrawsak & Street, 2006). When task demands confine performance to on-line processes, L2 speakers tend to have more difficulty with comprehension (Clahsen & Felser, 2006). Thus, comparing L1 and L2 performance under variable task demands can be instructive with respect to the role of centralized control processes in “good” versus merely “good-enough” thematic renderings of a sentence.

The Present Study

We first ran a single-task replication of Ferreira (2003) with native and non-native speakers of English to establish how grammatical voice and plausibility impacts performance for the less proficient group. If L2 performance differs from that of L1 due to factors unrelated to controlled processing effort—e.g., insensitivity to word-order cues and/or overreliance on semantic relations between sentence constituents—then voice and plausibility factors should interact with language group. Specifically, the impact of non-canonical order should have a smaller effect on the L2 group—specifically in the plausible passive condition—and/or the effect of semantic plausibility should have a larger effect on the L2 group—especially in the implausible active condition. However, if groups differ more in terms of the difficulty associated with processing an L2, then group differences, if any, should interact linearly with voice and plausibility—i.e., as a function of the difficulty associated with fully processing all the elements of the input.

We specifically recruited native speakers of French for the L2 group, because French active and passive structures are homologous to that of English (compare 2a to 2b). Both passives include an auxiliary verb—to-be/être—and both indicate the agentive noun with a prepositional phrase—by/par-NP. The similarity between English and French voicing makes it unlikely that comparative language variables per se should influence how the various cues are integrated by the L2 group. In the absence of direct evidence that L2 speakers are either less sensitive to linearity cues or more sensitive to semantic cues, any observed differences are likely a function of differences in processing fluency—specifically, the ease with which a veridical interpretation is generated. In Experiment 2, we address more directly the role of centralized control processes underlying group differences.

(2) a. The dog bit the man. / The dog was bitten by the man.

b. Le chien a mordu l’homme. / Le chien a été mordu par l’homme.

Experiment 1: Single Task Paradigm

Method

Participants

Participants were recruited from an Anglophone university in Montréal and from the surrounding community. Montréal is an English-French bilingual city, where French is dominant, but where many native English speakers live, and where large subsets of both Francophone and Anglophone populations attend schools, work, and socialize primarily in their native tongue. Prior to selection of participants, 308 individuals completed the Language Background Questionnaire (LBQ; Segalowitz & Frenkiel-Fishman, 2005), which measures respondents' self-reported language history and proficiency. Based on their responses to the LBQ, we selected 32 native speakers of English for the L1 group who met the following criteria: (1) They identified English as both their first and dominant language; (2) on a 5-point proficiency scale, ranging from 1 (no ability) to 5 (fluent ability), they rated their English proficiency as 5 for speaking, listening, and reading respectively; and (3) they reported no hearing, reading, or attention disability.

The L2 group consisted of 29 native speakers of French, who were selected from the same pool of LBQ respondents based on the following criteria: (1) French was identified as their first and dominant language; (2) English was identified as their L2, acquired only after the acquisition of French, and no earlier than the age of three (Mean Age of Acquisition=8.97, SD=4.3); (3) they reported using English at least 10% of the time in their daily lives; and (4) they reported no hearing, reading, or attention disability. Those participants who were psychology students at Concordia University were compensated with credit for their courses; otherwise, they were paid $10 CAD for their participation.

Table 2 presents age and language background characteristics for both groups, as well as independent samples t-tests and 95% confidence intervals to assess whether participants differed with regards to the relevant language background characteristics. Concerning English proficiency variables, the L2 group reported using English less often, Mean Difference (MD) =44.36%, and being less proficient at speaking, reading, and listening (MD =0.52) on a 5-point scale. Concerning French proficiency measures, the L2 group reported using French more often (MD=44.52%), and being more proficient at speaking, reading, and listening (MD=1.67). Although the L2 group's reported English proficiency levels were statistically lower than that of

Table 2

Means (Standard Deviations) of language background characteristics, independent-samples t-tests, and 95% confidence intervals.

	L1 (N=32)	L2 (N=29)	T-Test	95% CI (M_D)
Age	25.47 (7.9)	23.00 (5.3)	$t(59)=1.42$, $p=.16$	[-1.00, 5.94]
English Use (%)	88.84 (5.5)	44.48 (17.6)	$t(59)=13.81$, $p<.001$	[37.81, 50.91]
French Use (%)	10.22 (5.6)	54.74 (17.8)	$t(59)=-13.41$, $p<.001$	[-51.16, -37.89]
L3 Use (%)	0.94 (2.2)	0.78 (2.2)	$t(59)=0.29$, $p=.77$	[-0.97, 1.29]
English Fluency[a]	5.00 (0.0)	4.48 (0.6)	$t(59)=4.73$, $p<.001$	[0.29, 0.74]
French Fluency	3.28 (0.7)	4.95 (0.2)	$t(59)=-12.21$, $p<.001$	[-1.95, -1.39]

[a] T-tests for samples with unequal variances were based on the Welch degrees of freedom approximation.

the L1 group, they were nevertheless quite high. Mean ratings for speaking, reading, and listening fell between 4 (very good) and 5 (fluent ability). Thus, our L2 group may be best qualified as proficient late learners of English.

Materials

The stimuli consisted of 24 English item frames developed by Ferreira (2003), each with a simple transitive verb and two nouns (e.g., *bite_dog_man*). These frames were manipulated by voice (active, passive) and plausibility (plausible, implausible) to create 98 unique sentences (see Table 1 for an example). The sentences were divided into four orthogonal lists, consisting of six sentences from each of the four conditions and a set of 59 filler sentences that were common to all lists. Participants were assigned to lists in sequential order to assure that sentences were equally distributed throughout the sample. All sentences were read by the same female native speaker of English using natural speed and prosody, and were recorded for aural presentation during the experiment, at which time they were presented randomized through a pair of noise-cancelling headphones.

Procedure

Participants were tested individually using an automated paradigm with PsyScope X (Cohen, MacWhinney, Flatt, & Provost, 1993), modeled after Ferreira's (2003) procedure, with two exceptions. Firstly, we included only a subset of Ferreira's (2003) materials (24 of 72)—as our experiment addressed a more focused question—and we included fewer filler sentences (59 as opposed to 144). Secondly, we omitted comprehension questions regarding the patient of the sentence to increase the number of observations per cell.

Each trial began with the instruction *Press the black or white button to continue* displayed onscreen, followed by aural presentation of the sentence, which in turn was followed by a visually displayed prompt used to elicit sentence interpretations (see Figure 1). For the experimental trials, DOER elicited the agentive noun. For the filler sentences, three alternative prompts were used: (1) ACTION; (2) HOW; and (3) WHEN, which respectively called for an action/verb (e.g., baking), manner/adverb (e.g., quickly), or temporal information (e.g., last night) presented in the preceding sentence. Participants provided an oral response following the presentation of each prompt.

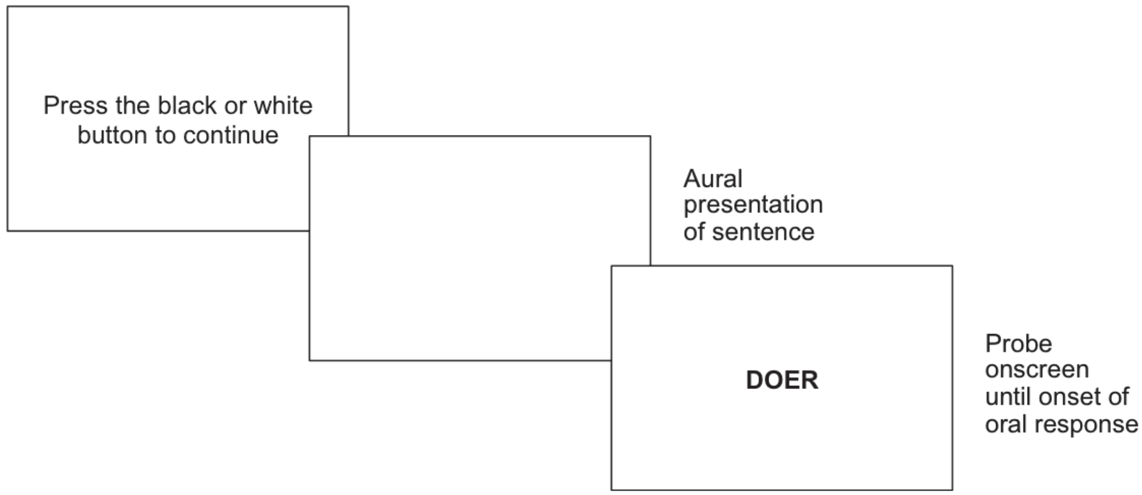

Figure 1. Trial sequence of Experiment 1.

During the instruction period, the experimenter defined each prompt and provided an example to test the participants' comprehension of the task. The agent task was described as follows: "DOER refers to the noun in the sentence that is doing the action. For example, in the sentence *The boy kissed the girl,* who is the DOER?" The experimenter used a similar script to introduce each of the filler probes. The session then proceeded with eight randomized practice trials with two trials per probe type. For the experimental trials, one sentence was presented in the active voice, *The butler retrieved the wine,* and one in the passive voice, *Mars was inhabited by aliens last December.* If participants made an error during the practice trials, the experimenter prompted them using the following sample script: "Earlier you heard the sentence *Mars was inhabited by aliens last December.* What would be considered the DOER in that sentence?" The test sentence embedded within this script varied according to the sentence that was misinterpreted. If the participant did not spontaneously correct their mistake upon a second prompting, their data was excluded from analysis.

Results & Discussion

All participants qualified for data inclusion based on their performance during the practice trials. However, we excluded seven observations (0.48%) for trials in which participants failed to provide a task-relevant response—i.e., either the agent or patient noun. This occurred once in the L1 group and six times in the L2 group.

We fitted to the data a binomial logit mixed-effects model using the 'lme4' package (Bates, Maechler, & Bolker, 2013) in R (R Core Development Team, 2012). Accuracy was regressed on voice, plausibility, language group, and all possible first-order interaction terms with participants and items entered as random effects. Compared to a null model consisting of only random predictors, the full model provided a better fit to the data, $\chi^2(6)=178.16$, $p < 0.001$. A summary of fixed effects is presented in Table 4, and descriptive statistics—calculated with data averaged by participants—are presented in Figure 2.

Voice produced the largest effect on response accuracy, with participants reporting the patient noun as the agent of passive sentences 19.23% of the time. Plausibility also affected response accuracy, but to a lesser degree as implausible sentences were misinterpreted 13.33% of the time. The interaction of voice and plausibility was not statistically significant. As in previous studies, implausible passives produced the highest error rate, 25.28%, of all sentence types. The effects of voice and plausibility were present in both L1 and L2 groups, with no significant

Table 3

Logistic regression of response accuracy on (A) Voice, (B) Plausibility, (C) Language group and all first order interaction terms.

	Estimate	SE	z	*p*	*OR* (e^{E})	95% CI [OR]
Intercept	1.52	0.30	5.06	0.000	4.57	[2.54, 8.23]
(A) Voice	3.87	0.65	5.96	0.000	48.04	[13.45, 171.52]
(B) Plausibility	1.18	0.32	3.66	0.000	3.24	[1.73, 6.08]
A * B	-1.27	0.65	-1.95	0.052	0.28	[0.08, 1.01]
(C) Language	-0.34	0.39	-0.86	0.391	0.71	[0.33, 1.54]
A * C	-0.42	0.69	-0.61	0.540	0.66	[0.17, 2.53]
B * C	-0.32	0.42	-0.76	0.445	0.73	[0.32, 1.65]

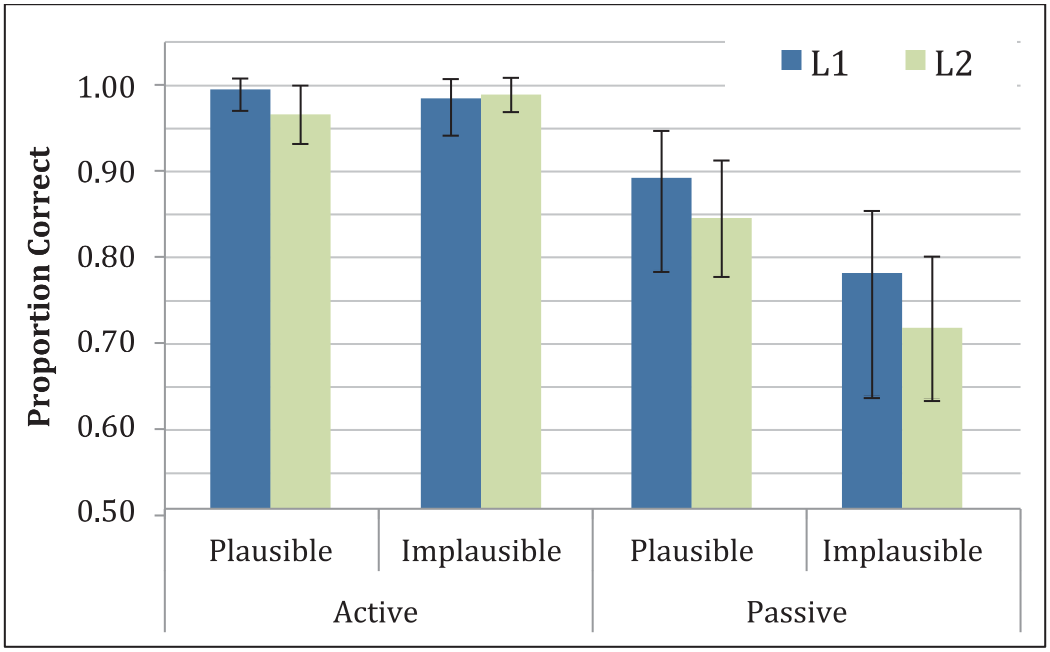

Figure 2. Proportion correct responses for all sentences for L1 and L2 speakers of English. Error bars represent the standard error of the mean.

differences between them (see Figure 2) and no interactions between the within-subjects linguistic factors and language group.

The results suggest that the processes underlying “good-enough” interpretation errors are similar in native and proficient non-native speakers of English. Insofar as the task measures grammatical knowledge, the results suggest that both groups had sufficient knowledge of the grammar to assign thematic roles correctly in all conditions most of the time. And both groups produced more errors with non-canonical structures, particularly when the sentences were implausible. The pattern of effects suggests that proficient non-native speakers of English—whose L1 is French—are not less sensitive to distributional probabilities of English and are not more biased by semantic plausibility. The main conclusion to be drawn from these data is that Ferreira’s (2003) misinterpretation effect is robust and replicable in both native and proficient L2 speakers. The data, however, do not speak to the question of whether L2 success was achieved with greater controlled processing effort and/or the extent to which domain-general resources were required in both groups to generate a veridical response. We turn now to a dual task experiment designed to investigate further the degree to which good-enough errors are affected by, or interact with non-linguistic task demands as a means of better operationalizing the factors that modulate misinterpretation.

Experiment 2: Dual Task Paradigm

The motivation for this experiment is to investigate the role of domain-general processing in veridical responding. If off-line thematic assignment requires the engagement of non-linguistic control process, success in L1 and L2 speakers should interact with performance on a secondary working-memory (WM) task. For instance, participants may have to inhibit linearity cues for agency, and/or monitor syntactic markers off-line when coordinating a response to the DOER probe. To the extent that challenges like these underlie good-enough misinterpretations, performance on the thematic task should trade-off with performance on the WM task.

In this experiment, participants were presented with six randomly generated digits prior to sentence onset, and recall was cued following their report of the agentive noun. Thus, participants were required to maintain the digits in WM throughout the thematic assignment task. Under these conditions, comprehension outcomes may be a function of at least two factors: (1) the extent to which domain-general control processes are required to generate and report a veridical response; and (2) the interpreter’s motivation to engage those processes at the expense of performance on the WM task. L1 and L2 speakers may differ with respect to both factors. Firstly, given the L1-L2 fluency gap, L2 comprehension may be relatively more dependent on the control system during structural processing (Paradis, 2004; 2009; Segalowitz, 2010; Ullman, 2001; 2015). And perhaps consequently, L2 speakers may allocate more attention and effort to language processing tasks when confronting demands from non-linguistic information sources. Thus, at least two possible outcomes could demonstrate a role for domain-general processes in veridical responding. If both groups distribute effort across tasks equally, then, the WM load should interfere with both L1 and L2 performance, but to a greater degree for the L2 group under the assumption that generating an interpretation requires more controlled processing in non-native speakers. Alternatively, if language groups distribute effort differently across tasks, group differences in comprehension should show evidence of task trade-offs. The group that performs better on the language task should perform worse on the digit task, indicating that interpretation success draws on processing resources that support digit maintenance and recall.

Method

Participants

A unique group of individuals—that did not participate in Experiment 1—were selected from the same pool of LBQ respondents as Experiment 1 following the same inclusion criteria. The mean age of English acquisition for the L2 group was 9.29, *SD*=4.2. Table 4 summarizes the relevant characteristics of both groups, with independent-samples t-tests and 95% confidence intervals. As in Experiment 1, the L2 group reported using English less often (M_D=44.36%), and being less proficient at speaking, reading, and listening (M_D=0.52) than the L1 group, whereas they reported using French more often (M_D=44.52%), and being more proficient at speaking, reading, and listening in French (M_D=1.67) than the L1 group. Note that although we report a statistically significant difference in age between groups (L2 > L1), we do not regard this as substantively significant as both groups were comprised primarily of adults in their twenties.

Dual Task Procedure

All aspects of the sentence interpretation task, including materials, were the same as in Experiment 1, with the exception that an extraneous working memory load was added as follows. At trial onset, a fixation mark was displayed for 1000 milliseconds (ms) followed by six randomly generated digits presented visually in succession. Each digit appeared for 600 ms followed by an inter-stimulus interval (ISI) of 1000 ms. At the offset of the sixth ISI, the

Table 4

Means (Standard Deviations) of language background characteristics, independent-samples t-tests, and 95% confidence intervals.

	L1 (N=29)	L2 (N=29)	T-Test	95% CI (M_D)
Age	20.67 (4.6)	25.42 (7.9)	$t(58)=-2.81$, $p=.007$	[-8.15, -1.37]
English Use (%)	86.72 (9.3)	40.81 (16.5)	$t(58)=13.15$, $p<.001$	[38.93, 52.91]
French Use (%)	10.79 (7.9)	56.10 (17.7)	$t(58)=-12.60$, $p<.001$	[-52.50, -38.11]
L3 Use (%)	2.48 (4.1)	3.10 (5.3)	$t(58)=-0.54$, $p=.62$	[-3.05, 1.82]
English Fluency[a]	5.00 (0.0)	4.34 (0.6)	$t(58)=6.36$, $p<.001$	[0.45, 0.87]
French Fluency[a]	3.63 (0.8)	5.00 (0.0)	$t(58)=-9.99$, $p<.001$	[-1.64, -1.09]

[a] T-tests for samples with unequal variances were based on the Welch degrees of freedom approximation.

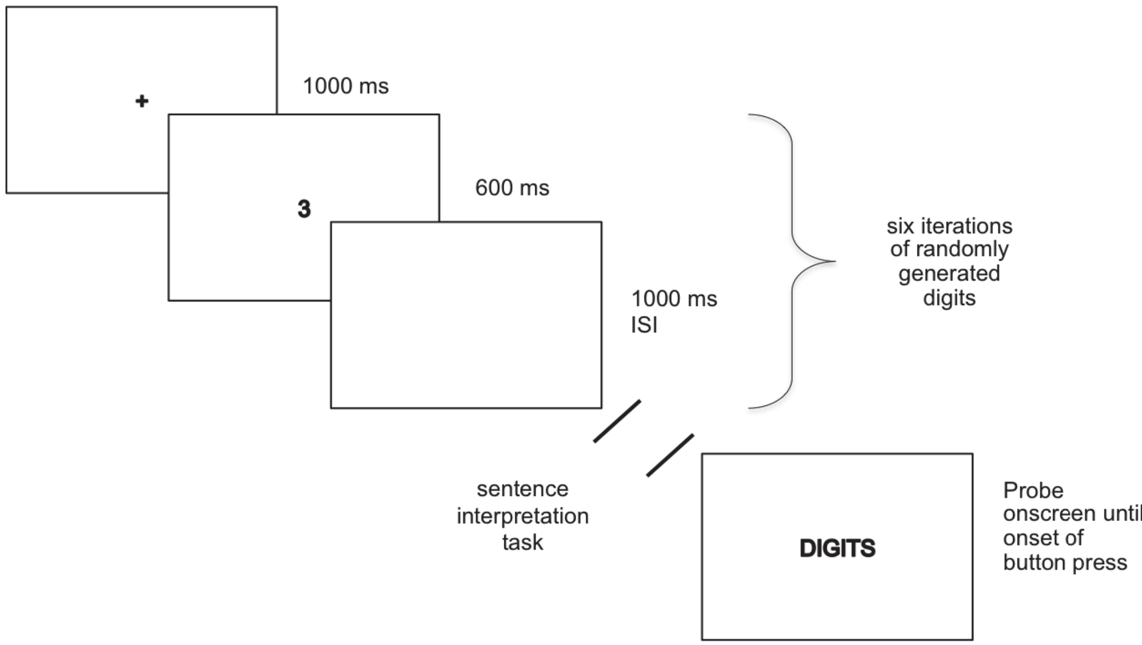

Figure 3. Trial sequence in Experiment 2. Duration is presented in milliseconds (ms).

sentence comprehension task progressed as in the single task condition, after which the prompt DIGIT appeared onscreen. Participants recited as many digits as they could recall from the beginning of the trial, and then pressed a button to proceed to the next trial (see Figure 3).

Results & Discussion

Language Comprehension Task. All participants qualified for data inclusion based on their performance during the practice trials. However, we excluded 18 observations (1.29% of the dataset) for trials in which participants failed to provide a task-relevant response—i.e., either the agent or patient noun. This occurred nine times in each group.

We fitted to the data a binomial logit mixed-effects model using the 'lme4' package (Bates et al., 2013) in R (R Core Development Team, 2012). Accuracy was regressed on voice, plausibility, language group, and all possible first-order interaction terms with participants and items entered as random effects. Compared to a null model consisting of only random predictors, the full model provided a better fit to the data, $\chi^2(6)=217.01$, $p < 0.001$. A summary of fixed effects is presented in Table 5, and descriptive statistics—calculated with data averaged by participants—are presented in Figure 4.

The fitted model again replicated the effects of voice and plausibility on thematic assignment, with increased errors for passives (22.75%) compared to actives (1.89%), and for implausible sentences (16.31%) compared to plausible sentences (10.21%), but no interaction between these factors. Unlike Experiment 1, our dual task procedure generated group effects, with more passive errors in the L1 group (27.74%) than the L2 group (21.53%). As will be described next, these group differences traded-off with performance on digit recall, thus, we will interpret the results of the two tasks in tandem.

Digit Recall Task. Within each trial, digit responses were assigned a serial position from 1 to 6 and each position was scored on a scale from 0-2. Utterances in the $6 + n^{th}$ position were treated as commission errors. We assigned 2 points to digits recalled in the correct position, 1 point for digits recalled in the incorrect position, and 0 points for omitted or committed digits. Thus, digit recall scores ranged from 0-12. Descriptive statistics are presented in Figure 5 as percentage scores.

Digit recall data was fitted to a linear mixed-effects model in the 'lme4' package (Bates et al., 2013) in R (R Core Development Team, 2012). Digits scores were regressed on voice, plausibility, language group, and all first-order interaction terms with participants and items

Table 5

Logistic regression of response accuracy on (A) Voice, (B) Plausibility, (C) Language group and all first order interaction terms.

	Estimate	SE	z	*p*	*OR* (e^{E})	95% CI [OR]
Intercept	0.82	0.28	2.94	0.003	2.27	[1.31, 3.93]
(A) Voice	3.46	0.50	6.91	0.000	31.93	[11.96, 85.25]
(B) Plausibility	0.88	0.27	3.25	0.001	2.41	[1.42, 4.1]
A * B	0.61	0.69	0.88	0.379	1.84	[0.47, 7.14]
(C) Language	0.55	0.27	2.04	0.041	1.74	[1.02, 2.95]
A * C	-0.88	0.62	-1.43	0.154	0.41	[0.12, 1.39]
B * C	-0.35	0.38	-0.91	0.361	0.71	[0.33, 1.49]

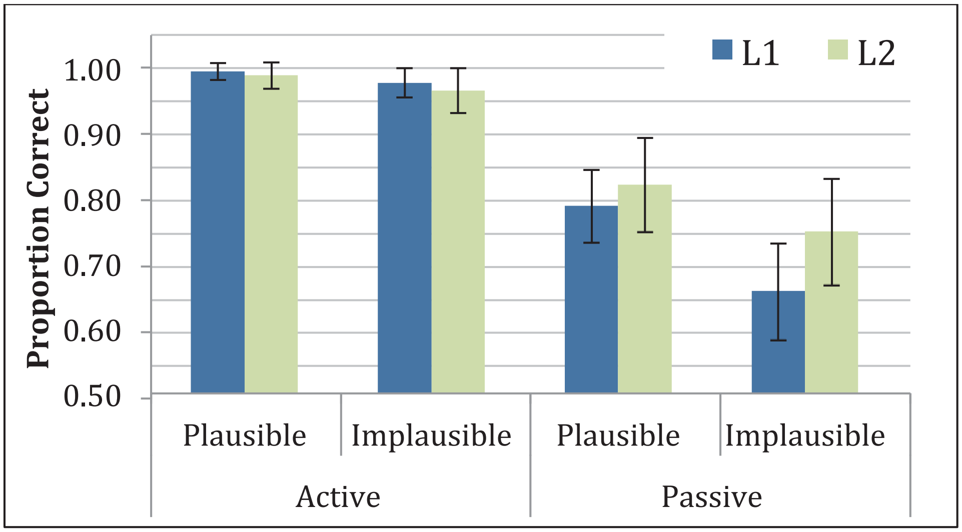

Figure 4. Proportion correct responses for all sentences for L1 and L2 speakers of English. Error bars represent the standard error of the mean.

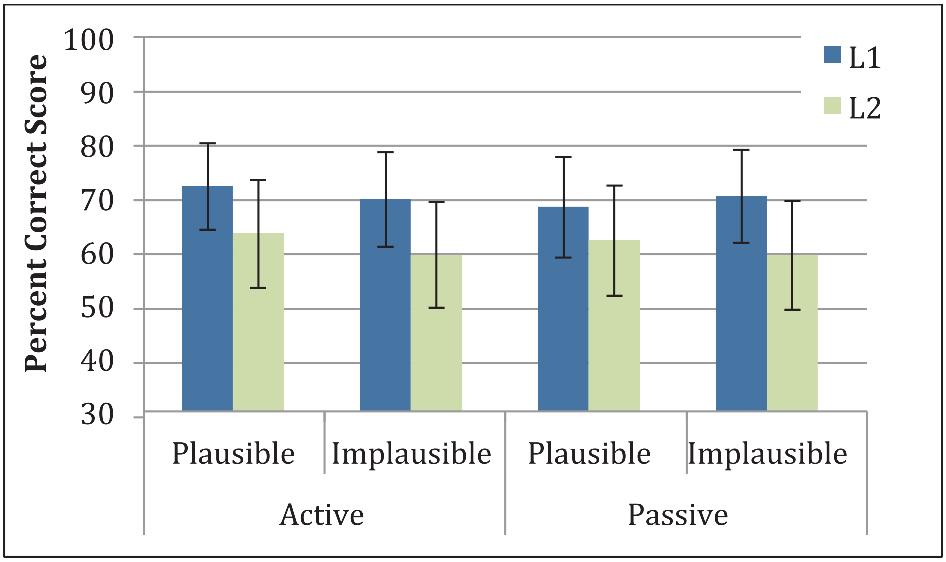

Figure 5. Mean percent digit recall scores as a function of voice, plausibility, and language. Error bars represent 95% confidence intervals with *n* equal to the number of participants.

entered as random effects. Compared to a null model consisting of only random predictors, the full model provided a better fit to the data, $\chi^2(6)=81.51$, $p < 0.001$. A summary of fixed effects is presented in Table 6.

Digits scores did not interact with either of the within-subjects linguistic factors (voice, plausibility), suggesting that domain-specific sources of processing difficulty did not interfere with digit maintenance in either group. Differences did emerge, however, as a function of language background, with the L1 group performing better than the L2 group consistently in each of the sentence conditions (see Figure 5).

The combined observations of language comprehension and digit performance illustrate that the relative success of L2 comprehension was obtained at the expense of digit recall. Similarly, the relative success of L1 digit recall came at the cost of accuracy on the thematic assignment task. These trade-offs imply that the cognitive resources that underlie digit recall could also be used when rendering thematic roles of non-canonical sentences in both native and non-native speakers. Successful interpretation on Ferreira's (2003) thematic assignment task thus requires some measure of controlled processing effort.

Concerning L1-L2 differences, in particular, our data speak more directly to differences in attention allocation than differences in language processing. Our non-native participants were more successful than their native speaking counterparts primarily because they distributed more effort to the language comprehension task. Although their tendency to prioritize language over mnemonic task goals may stem from differences in language processing—i.e., the need to engage in more controlled processing effort when interpreting sentences—the presence of different task selection strategies in our L1 and L2 participants prevents us from casting comprehension outcomes in terms of different interpretation mechanisms per se. We can conclude, however, that the dynamics of veridical versus good-enough responding in L1 and L2 speakers is determined to some extent by the interpreter's motivation to engage in effortful control processes when coordinating a response to the experimental probe. The pattern of results suggests that L1 participants were more inclined to default to a good-enough rendering of the sentence, whereas the L2 participants worked harder to generate a veridical response. Critically, group differences are not likely to have been associated with differential sensitivities to linearity and semantic agency cues for a number of reasons. Firstly, French active and passive sentences are homologous English actives and passives; thus, native-language parsing

Table 6

Linear regression of digit recall scores on (A) Voice, (B) Plausibility, (C) Language group and all first order interaction terms.

	Estimate	SE	t	*p*	95% CI [OR]
Intercept	8.58	0.28	30.20	0.000	[8.03, 9.14]
(A) Voice	0.12	0.27	0.43	0.678	[-0.41, 0.64]
(B) Plausibility	-0.10	0.27	-0.39	1.241	[-0.63, 0.42]
A * B	0.31	0.31	1.01	0.318	[-0.29, 0.91]
(C) Language	-1.66	0.28	-5.96	0.000	[-2.21, -1.12]
A * C	-0.10	0.31	-0.34	1.214	[-0.71, 0.5]
B * C	0.35	0.31	1.14	0.257	[-0.25, 0.96]

preferences are likely to have been similar in our Francophone and Anglophone participants. Secondly, baseline patterns of performance obtained in our single task experiment showed equivalent sensitivity to voice and plausibility in both groups. Finally and most importantly, the pattern of effects observed in Experiment 2 again showed no interaction of voice and plausibility with language group. Non-native speakers were clearly sensitive to linearity cues as they misinterpreted passive sentences 22% of the time (even more frequently than Ferreira's (2003) participants, who misinterpreted passives 19% of the time). Similarly, non-native speakers were not more biased by semantic cues, as they misconstrued implausible sentences less frequently than our native speaking participants (14% and 18% respectively). Group differences thus emerged primarily as a function of task selection; in particular, the tendency for native and non-native speakers to allocate effort to linguistic and mnemonic outcomes differently. In the General Discussion, we will elaborate on the implications of our results for the nature of sentence processing and, in particular, the good-enough model.

Discussion

Misinterpretation of passive sentences has served as one of the chief lines of evidence for the hypothesis that the language comprehension system systematically underspecifies structure during processing (Christianson, 2016; Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007; Karimi & Ferreira, 2015). Although previous studies have shown that misinterpretations occur only in a minority of instances—10-12% of plausible passives and 26% of implausible passives (Christianson et al., 2010; Ferreira, 2003)—these errors were frequent enough to illustrate that extra-grammatical variables influence off-line comprehension outcomes. What was less clear from previous studies was whether the response criterion—i.e., participants' oral report of the agentive noun—reflected the default output of the parsing mechanism, or whether a non-trivial proportion of variance could be attributed to extra-linguistic task factors as well. In the set of experiments reported here, we demonstrated that indeed coordinating a veridical response to an off-line comprehension probe draws on generalized control resources.

There are several possibilities for the role that generalized controlled processes play in supporting interpretation: (1) controlled mechanisms may be required to complete a structural parse of non-canonical sentences; (2) controlled mechanisms may be required to adjudicate between competing cues for agency as a thematic rendering of the sentence develops; or (3) controlled mechanisms may be required to select the agent from the output of the parser after thematic roles have been faithfully rendered. Although we cannot definitively rule out any of these hypotheses, we propose that our task selection effect may be most consistent with the third explanation. In particular, the finding that non-native speakers outperform their native-speaking conspecifics may suggest that responses draw heavily on metalinguistic processes extraneous to parsing.

Consider that, in the course of semantic composition, comprehenders may not consciously track sentence constituents in terms of their grammatical, semantic, and thematic classes. Therefore, when asked to report the agent of a sentence, participants must generate a metalinguistic formulation of their interpretations. While such formulations are certainly informed by sentence interpretations, they provide only an indirect index of comprehension, because they also require a secondary stage of metalinguistic processing. This secondary stage may itself be vulnerable to heuristics even when interpretations are initially correct. In fact, there is some evidence that the passive misinterpretation effect may be a function of how comprehension is measured, rather than of comprehension *per se*. Specifically, Bader and Ming (2016; AMLaP Proceedings) reported that when interpretation is assessed in a speeded plausibility judgment task, participants provide correct judgments of plausible and implausible passives. This can occur only when participants correctly determine the relationships between lexical items, yielding a compositional meaning. Nevertheless, when asked to identify the thematic roles of these sentences in a subsequent stage of the task, participants who had initially judged the plausibility of sentences correctly sometimes confused thematic roles. This suggests that generating a metalinguistic conceptualization of thematic roles at sentence offset may be a taxing process in itself, and moreover, that this process may be decoupled from semantic composition.

If good-enough errors are indeed a function of metalinguistic processing rather than interpretation, the effect may occur because passive sentences topicalize their patients (Bock, 1986; Tolmin, 1983). More specifically, comprehenders may construe a sentence such as '*The dog was bitten by the man'* as being *about* a dog that was bitten. Passive topicalization may therefore place the patient noun in the spotlight of attention, with the details of the predicate—i.e., that it was a man who did the biting—fading into the periphery. Therefore, when reporting the agent of passive sentences, control processes may be recruited to inhibit the topic of the sentence in order to retrieve a less salient constituent.

What is Effortful About Processing Passive Sentences?

Whether misinterpretations reflect response interference during metalinguistic processing or underspecification of structure, our task selection effects suggest that task demands and their interactions with other variables modulate off-line responses, and are therefore critical to our understanding of good-enough processing dynamics—a view expressed equally by proponents of the good-enough framework (Christianson, 2016; Lim & Christianson, 2013a, b). Constraints associated with domain-specific and domain-general resources, individual differences, as well as response measurement are especially relevant to the operationalization of *effort* and in particular to our understanding of what is and what is not difficult about processing passive structures.

In the present experiments, we manipulated difficulty at three distinct levels of measurement: the linguistic signal, the language processing system, and extra-linguistic task demands. We demonstrated that the fidelity of off-line thematic assignment varies at each stratum of experimentation in a pattern that establishes a role for domain-general control processes and maintains the previously established role of domain-specific factors (Christianson et al., 2010; Ferreira, 2003). Perhaps the most surprising outcome was the direction of the controlled processing effects, whereby less proficient speakers were more accurate interpreters under dual-task conditions.

Here it is worthwhile to pause on the implications of this result for L2 research before returning to the matter of language processing effort more generally. Given the presence of L1-L2 task trade-offs, we take the most conservative position available: we eschew generalizing our L2 findings beyond the population of late learners of English, whose first and dominant language is French, and who have achieved a fairly high level of L2 proficiency based on frequent and regular use of their L2. Other populations of non-native speakers who may have different experiences with English and whose native language may entail a different set of parsing preferences (see McWhinney, Bates, & Kleigl, 1984) would not necessarily show the pattern of performance we observed in our samples. We also eschew casting L1-L2 differences in terms of differential sentence processing mechanisms, as we have outlined in our discussion of Experiment 2.

What we can say with relative conviction is that our L1 and L2 participants differed primarily with respect to how attention was allocated to language versus mnemonic task goals.

There may be any number of reasons for this difference, including: (1) digit maintenance is more difficult and/or aversive for non-native speakers; (2) language comprehension failures are more ego-dystonic for—and in everyday life perhaps more costly to—non-native speakers; (3) native and non-native speakers prioritize tasks differently in an attempt to compensate for relative weaknesses in different domains and optimize total performance (akin to the task selection effects observed in younger versus older adults: Li, Lindenberger, Freund, & Baltes, 2001); or (4) any combination of these variables with others that we have not considered.

Regardless of the motivational underpinnings of our L1-L2 task selection effects, the effect of domain-general constrains on interpretation is instructive with respect to the processes that contribute to off-line comprehension outcomes. Because the good-enough model treats these outcomes as the direct object of the language comprehension system (e.g., Karimi & Ferreira, 2015), it tends to localize effort to the syntactic substrate of the linguistic system. However, this assumption overlooks the intermediate role that the control system plays in linking the syntactic mechanism to the off-line response. Whether or not the parser *per se* should be subdivided into two language processing systems—one heuristic and one algorithmic—is ultimately a matter for future empirical research. But the current data do underscore the importance of considering another sort of a division of labour—one that dissociates the on-line comprehension mechanism from the centralized control system.

A variety of studies have shown that the control system participates in various aspects of sentence processing (Fedorenko & Thompson-Schill, 2014); for instance, maintenance of sentence constituents in WM (Just & Carpenter, 1992), conflict monitoring and resolution (Ye & Zhou, 2009), and reanalysis of misinterpretations (Hsu & Novic, 2016). Although there are certainly instances in which domain-general resources may be required to fully process the structure of a sentence initially (consider for instance the sentence in (5) below), in general the on-line parser has no difficulty outputting fully specified, structural renderings of the input (see Koornneef & Reuland, 2016). What seems most relevant to the question of processing difficulty is not whether syntax is algorithmic and dissociable from heuristic computations, but whether or not the structure-building strategies required to render a faithful interpretation of a phrase are automatized. When they are, we may also ask what are the non-structural strategies required to satisfy task demands and how do they mediate between the structural outputs and task outcomes.

(5) The teacher who the boy who the girl hugged tripped fell.

With respect to passive sentences, we propose that the mapping of roles to structure may be automatic, but *what is not* is the task of generating a metalinguistic conceptualization of the agentive role. We suggest that this challenge falls more naturally within the domain of the centralized control system. Thus, successful interpretation in our experiments was shown to vary with controlled processing effort, in particular, our non-native participants' motivation to generate a faithful rendering of the input at the expense of digit recall. An open question for future research is to determine whether this motivational difference is tied to different L1-versus-L2 language processing mechanisms. One intriguing possibility is that L2 speakers work harder on language comprehension tasks precisely because they have to—that is, their initial rendering of structure requires computation of metalinguistic representations, and in turn, more controlled processing effort. Addressing this question however will require controlling task selection options so that the distribution of attention across tasks is held constant for native and non-native speakers.

Conclusions

Native and proficient non-native speakers of English alike are sometimes inclined to misconstrue the agentive role of passive sentences, illustrating that when the ordering of thematic roles is non-canonical, interpreters have access to a superficial rendering of the input that conforms to distributional probabilities of the language. This non-veridical rendering may be a function of metalinguistic response interference, or the direct object of a superficial parse of the sentence. Either way, the tendency to misconstrue thematic roles is exacerbated when the semantics of passive structures diverge from the interpreter's intuitions about how the world works. The ultimate success of interpretation—as measured by the interpreter's oral report of the agentive noun—is impacted by domain-general constraints on information processing. Thus, the difficulty of generating a veridical response is determined to some extent by centralized control processes that mediate between the on-line parsing mechanism and the off-line response. These centralized processes may be required to inhibit pragmatic effects of passive topicalization and may very well be dissociated from the interpreter's ability to generate a veridical event representation. We propose that what is effortful about good-enough processing is not thematic assignment *per se*, but the need to generate a metalinguistic formulation of the output in the face of interference from pragmatic processes.

Chapter 3: Methodological Considerations for Manuscript 2

The current program of research was motivated by the hypothesis that the passive misinterpretation effect (Ferreira, 2003) reflects metalinguistic processing dynamics rather than parsing and interpretation. It was therefore hypothesized that L1-L2 differences in metalinguistic processing (Clahsen & Felser, 2006; Paradis, 2009) would modulate the effect. More specifically, because L2 comprehension processes are more dependent on metalinguistic representations (and thus engage in that form of analysis by default), using a second language may facilitate the metalinguistic analyses required to conceptualize the agent of a sentence. The previous chapter reported the results of two between-subjects experiments involving different groups of participants—one single task and one dual task—that provided partial support for this hypothesis. The single task experiment required participants to report the agents of active and passive sentences, and the dual task experiment required participants to perform the same task with the addition of a six-digit WM load. While there were no between-group differences in the single task experiment, the L2 group outperformed the L1 group on the language component of the dual task experiment. Crucially, this effect occurred as a function of attention allocation across tasks. The L1 group—who performed worse on the language task—performed better on the WM task. These trade-offs suggested that the process of conceptualizing the agent of a sentence depends on the control system. That is, thematic assignment success occurred at the expense of digit recall and vice-versa. The results also provided evidence that L1-versus-L2 status may modulate the passive misinterpretation effect.

However, the results did not provide more direct evidence for the hypothesis that L1-L2 differences in metalinguistic processing per se modulated the effect, and more importantly, that the errors might be dissociated from semantic compositional processes (in a native language). Perhaps the main limitation of the previous experiments was that they followed a between-subjects design involving four different samples, which may have confounded L1-L2 comparisons. One unintended group difference was that our L2 participants were on average more balanced bilinguals than our L1 participants. Specifically, our L2 group from Experiment 2 reported using their native language (French) 56% of the time, and their second language (English) 41% of the time. In contrast, our L1 group from Experiment 2 reported using their native language (English) 87% of the time and their second language (French) 11% of the time.

These differences in bilingualism may have been associated with systematic group differences in executive control (Bialystok, Craik, & Luk, 2012). Although there was no direct indication that this confound influenced our results—indeed when both language and mnemonic task outcomes were aggregated, performance was strikingly similar across groups, suggesting similar global levels of control—it may have nevertheless provided a source of bias that could be eliminated with a within-subjects design. Perhaps more importantly, because the overarching goal of this research is to investigate the mechanisms underlying L1-L2 differences in thematic assignment, vis-à-vis how they may dissociate comprehension processes from metalinguistic processes, a within-subjects design is necessary to reduce the likelihood of unequal attention allocation across L1 and L2 conditions.

Chapter 4 reports the results of a within-subjects experiment designed to address these limitations and test the hypotheses more directly. It was thought that a within-subjects design would reduce the likelihood of task selection effects, because participants would be exposed to all task conditions, and task instructions would be framed the same way in each condition. Therefore, it was assumed that measurement of an individual across levels of language and load would stabilize the general mindset and motivations that participants bring to the different conditions of the experiment (in contrast to a between-subjects design). The within-subjects design also eliminated any potential between-group confounds, including but not limited to degree of bilingualism, language proficiency, and WM capacity (WMC). Participants also completed a secondary WMC task (Conway et al., 2005) to measure how performance across conditions might be related to generalized control processes, and reaction times (RT) were recorded and used to further test the hypotheses.

Other small changes were made to the task to enhance measurement validity: (1) digits were presented lexically (THREE/TROIS) rather than numerically to control for language of activation, and in turn, minimize potential language switching costs within trials; (2) the inter-stimulus interval between digits was decreased from 1000 ms to 400 ms to prevent rehearsal prior to sentence onset and to create potentially greater interference for sentence comprehension processes. As will be shown in the subsequent chapter, these methodological changes achieved the desired outcome by eliminating between-language task selection effects and providing support for the hypotheses in the predicted direction. Specifically, we found a task-by-language crossover interaction, such that bilinguals were more successful assigning agents in their L2 during single task blocks, but were more successful in their L1 during dual task blocks. This pattern suggested that L2 metalinguistic comprehension processes indeed facilitated access to a conceptualization of the agentive noun but only when the control system—on which metalinguistic processes depend—was permitted to operate freely. The pattern suggested moreover that L1 comprehension processes may be dissociated from this metalinguistic component of the task, as L1 performance remained stable in the dual task condition. Reaction times and WMC associations in the single task condition also provided evidence for greater linguistic-to-metalinguistic switch costs in L1 than L2, suggesting that L2 agentive representations are derived from the same essential metalinguistic control processes that support comprehension, whereas L1 agentive representations require switching from semantic compositional to metalinguistic systems of analysis.

Chapter 4: Metalinguistic Switch-Costs Drive "Good-Enough" Misinterpretations of Passives

When interpreting natural speech, we are rarely required to think about how incoming sentences are structured and map onto meaning. Generally, semantic composition—the process by which we assign a correct interpretation to a sentence based on lexical denotations and syntactic structure—ensues automatically, without the need for effort, intention, or conscious deliberation (Fodor & Lepore, 2002). However, several experiments appear to challenge the notion that sentence comprehension is an effortless process, by showing that interpreters sometimes report incorrect interpretations of linguistic inputs (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Christianson, Luke, & Ferreira; 2010; Ferreira, 2003; Frazier & Clifton, 2015). Building on a number of such observations, the good-enough approach to language comprehension suggests that the processes required to fully integrate the structure and meaning of a sentence are effortful and may therefore be truncated prematurely if a "good enough" (though potentially incorrect) interpretation is available (Karimi & Ferreira, 2015).

One of the main lines of evidence for the good-enough approach comes from the apparent misinterpretation of implausible passive sentences, such as *The dog was bitten by the man.* Ferreira (2003) showed that when participants were asked to report the agent of such sentences, they blundered 26% of the time, stating for instance that *the dog* was the agent of the event—the one doing the biting. According to Ferriera and colleagues (Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007; Karimi & Ferreira, 2015), this thematic assignment error suggests that the language comprehension system has difficulty fully processing the structural elements of the input, and therefore relies on a set of "fast and frugal", but error-prone, heuristics for determining the meaning of a sentence.

In the present study, we test the good-enough approach against an alternative hypothesis. We propose that the semantic compositional processes underlying the interpretation of passive sentences are automatic, and the resulting interpretations are correct. However, errors occur as a function of output interference, when participants attempt to generate a metalinguistic conceptualization of the agentive noun. Before we outline our hypothesis and approach in detail, we briefly review the good-enough account of the passive misinterpretation effect.

The Passive Misinterpretation Effect: The Good-Enough Approach and Alternative Hypotheses

In Ferreira's (2003) paradigm, participants listened to active and passive sentences that were either plausible or implausible (see Table 7). Participants were then cued at sentence offset to report the agent (or patient) noun of these sentences—that is, the doer (and undergoer) of the events described by the sentences. Ferreira (2003) found that participants reported the agentive noun correctly most of the time; however, whereas they performed at ceiling for active sentences in both their plausible and implausible variants, they misconstrued the agents of plausible passives 12% of the time, and of implausible passives 26% of the time.

Ferreira (2003) proposed that because interpretations were correct in the majority of instances for all sentence conditions, participants had the requisite grammatical knowledge to correctly interpret active and passive sentences alike. However, the fact that they erred some of the time suggested that they sometimes used heuristics, which interfered with the structurally licensed interpretation. One proposed heuristic concerns the linear ordering of sentence constituents (Bever, 1970). In English, agent and patient roles typically map onto the first and second nouns of a sentence respectively, providing a superficial cue for determining who did what to whom. Thus, to the extant that interpreters rely on word order to determine how the arguments of a verb are interrelated, interpretations will go awry for passives, which bear a non-canonical *patient-verb-agent* sequence. Another relevant heuristic, namely the plausibility strategy, is used when interpreters attempt to combine lexical constituents in the most plausible way independent of the structure (Bever, 1970), resulting in interpretations of implausible sentences that conform to regularities in the world but violate the meaning of the input.

As can be seen in Table 7, Ferreira's (2003) results provide evidence that linguistic inputs activate both linear order and plausibility heuristics, which bias interpreters' off-line judgments (see also Christianson et al., 2010; Riven, de Almeida, Segalowitz, in preparation). Thematic assignment errors are found primarily for passive sentences—when the ordering of agents and patients are non-canonical—and this effect is exacerbated when the passive is implausible. But the precise locus and mechanism of the error—that is, whether it occurs as a direct function of semantic compositional processes—is not entirely clear. According to the good-enough approach, syntactic processing is effortful and slow; and therefore, interference is localized to the

Table 7

Sample sentences used by Ferreira (2003) and percent correct report of the agentive noun

Condition	Sentence (Agent)	Accuracy
Active, Plausible	The dog bit the man.	99%
Active, Implausible	The man bit the dog.	99%
Passive, Plausible	The man was bitten by the dog.	88%
Passive, Implausible	The dog was bitten by the man.	74%

syntactic substrate of the parser: Errors occur because the language comprehension system curtails syntactic processing if it is satisfied with the heuristic interpretation (Karimi & Ferreira, 2015). We propose however that the difficulty of reporting agents of passive sentences may arise only after semantic composition, when participants attempt to coordinate a response.

Consider that in the natural course of language processing, interpreters do not consciously track elements of the input in terms of their grammatical or thematic classes. This kind of information is mere metadata to the outputs of comprehension processes. Thus in order to report the agent of a sentence, the interpreter must initiate a post hoc metalinguistic formulation of the output. While this analysis is certainly informed by interpretations, the response generation process per se occurs outside the realm of sentence processing. We hypothesize more specifically that the executive control system is recruited for this metalinguistic stage of processing (Riven, de Almeida, & Segalowitz, in preparation), and errors occur because attention is biased by pragmatic factors—including linearity and plausibility cues—even when comprehension processes are faithful to the input.

The Present Study

The goal of this experiment is to test the hypotheses that (a) off-line thematic assignment entails metalinguistic thinking, which can be dissociated from parsing and interpretation mechanisms, and (b) linearity and plausibility heuristics interfere primarily with controlled metalinguistic processes. We tested our hypotheses in a single- versus dual-task experiment with Anglophone bilinguals, who acquired French as a second language later in life. The rationale of our design is as follows: first and second language (henceforth, L1 and L2 respectively) processing is marked by a fluency gap that affects the automaticity of language comprehension processes (Segalowitz, 2010). Because an L2 differs from an L1 in terms of age of acquisition, quantity and quality of exposure, circumstances of learning, and regularity of usage, differences emerge in how the underlying language representations are stored and used (Paradis, 2004; 2009; Ullman, 2001). The most relevant difference concerns the degree to which parsing and interpretation relies on implicit semantic composition strategies (L1>L2) or explicit metalinguistic processes (L2>L1).

Metalinguistic processing refers to any cognitive analysis that draws on what the interpreter consciously believes about the norms and rules of language. While both L1 and L2 speakers may use metalinguistic knowledge to some extent depending on the circumstances, L2 speakers are more dependent on metalinguistic processing for comprehension. This is partly because L2 acquisition entails explicit memorization of grammatical rules, and partly because most L2 speakers never fully automatize the implicit semantic composition strategies that drive L1 parsing and interpretation (Paradis, 2009). These L1-L2 differences affect how the control system is engaged in language interpretation (Paradis, 2009). Implicit semantic composition processes allow native speakers to automatically integrate structural information on-line (Clahsen & Felser, 2006), but may also entail delays or costs in engaging the control system when metalinguistic thinking is required. In contrast, the relative absence of implicit processing strategies in L2 necessitates higher initial engagement of the control system during L2 interpretation, potentially easing access to metalinguistic representations when deliberation is permitted, but impeding interpretation when it is not (Clahsen & Felser, 2006).

In the single task condition, bilinguals listened to active and passive sentences that were either plausible or implausible, and were cued at sentence offset to report the agentive noun. In the dual-task condition, bilinguals performed the same task with the addition of a six-digit working-memory (WM) load. It was hypothesized that single-task performance would provide evidence for a *metalinguistic buffer effect*—specifically, bilinguals would perform better in their L2 due to increased access to metalinguistic knowledge during interpretation. However, because metalinguistic processing is dependent on the control system, it was further hypothesized that patterns of L1-L2 performance would be reversed in the dual task condition. That is, L1 interpretations would persevere to a greater degree under load given the availability of automatic semantic composition processes. To the extent that such effects were to be obtained, the data would provide evidence that (a) L1 compositional processes are automatic, rather than effortful and slow, (b) off-line thematic assignment is an index of both linguistic and metalinguistic processes, and (c) thematic assignment errors may occur primarily as a function of metalinguistic processing dynamics. That is, what is effortful about reporting the agents of passive sentences is that it requires metalinguistic formulations, which may be skewed by how passive sentences draw attention to event constituents.

Method

Participants

Participants were recruited from the psychology student population at Concordia University in Montréal—a city where French is the dominant language, but where a large number of Anglophones live and attend schools in their native tongue. For the purposes of this experiment, we aimed to recruit Anglophones with sufficient knowledge of French to be able to interpret passive structures, but for whom there is a measurable fluency gap between their L1 (English) and their L2 (French). Based on self-report data, we recruited students with the following inclusion criteria: (1) English was their first and dominant language from birth; (2) French learning was initiated subsequent to English acquisition; (3) relative frequency of English usage was at least 50% of current interactions and at least 15 percent points greater than French and any other language; (4) subjective ratings of English proficiency, aggregated across independent measures of reading, listening, and speaking, were no less than 4.5/5 and each constituent measure was greater than the corresponding French-proficiency rating.

For convenience, students were initially screened electronically using an 8-item questionnaire—posted on a campus-wide intranet site—designed to determine gross-level suitability for participation, vis-à-vis that English was their L1 and French was their L2. Once in the lab, participants completed a more detailed form, adapted from the Language Background Questionnaire (LBQ; Segalowitz & Frankiel-Fishman, 2005), administered following participation in the experimental task. Eighteen of 68 students who initially met gross-level criteria for participation were later excluded based on their responses to the LBQ. The most common reason for exclusion was near-perfect bilingualism, which neutralizes the fluency gap required for the experiment. Other reasons for exclusion were that English was not a true L1—for instance, there was significant exposure at home to a subdominant "third" language prior to English acquisition—or, in a similar fashion, there was measurable exposure to French in early childhood overlapping with L1-English acquisition. An additional five participants were excluded from analysis because they failed to demonstrate sufficient linguistic knowledge during practice to qualify for data inclusion (specified in more detail below). Among our final sample of 45 English-French bilinguals, the average age of French acquisition was 7.64, SE=1.03, 99% CI [5.00, 10.29], and an L1-L2 fluency-gap was reported in terms of relative frequency of usage (%)

Table 8

Means [SD] of L1 and L2 percent usage and fluency rating scores, as well as L1-L2 mean differences (M_D), M_D effect sizes, and 99% confidence intervals around M_D.

	Percent Usage	Fluency Rating
L1 (English)	81.47% [12.48]	4.98 [0.08]
L2 (French)	15.73% [11.67]	3.66 [0.72]
M_D [*SE*]	65.73 [3.52]	1.32 [0.11]
Hedge's *g*	18.53	2.04
99% CI	[56.66, 74.8]	[1.05, 1.59]

and self-rated fluency on a 5-point scale. Table 8 presents descriptive statistics and mean-difference (M_D) estimation metrics for these variables.

Materials & Design

The stimulus set consisted of 24 English sentence frames developed by Ferreira (2003), each with a transitive verb and two nouns, forming a basic argument structure frame such as *[bite [man, dog]*. These 24 frames were manipulated by voice (active, passive) and plausibility (plausible, implausible) to create 96 unique sentences, which were then translated into French.

Norming of French Translations. One of Ferreira's (2003) items, *The hunter shot the deer,* was changed because there was no satisfactory homologous French translation. For the French version of this sentence, we used instead *Le chasseur a peiegé le lapin*, meaning *The hunter trapped the rabbit.* All 96 translations were screened for grammatical errors by 25 native French speakers. Following a second phase of translation, in which initial grammatical errors were corrected, an additional 28 native French speakers provided plausibility ratings for active sentences in both their plausible and implausible framings. Mean ratings of plausible actives on a 7-point scale was 6.36, SE=0.347, 95% CI [5.70, 7.06], and mean ratings of *implausible* actives was 2.46, SE=0.481, 95% CI [1.52, 3.41], with a mean-difference effect size of d=2.69. Note that plausibility ratings for English sentences, as reported by Ferreira (2003), were 5.82, SD=0.93, and 2.08, SD=1.00 for plausible and implausible actives respectively. The full set of Ferreira's (2003) materials and our French translations are presented in Appendix A. All sentences were read by a native speaker of each respective language, using natural speed and prosody, and were recorded for aural presentation during the experiment.

Material Lists. The 96 English and 96 French sentences were equally divided into eight counterbalanced lists, consisting of three unique sentences per sentential condition. These 16 lists were distributed into eight semi-counterbalanced groups, to which participants were assigned. Each group contained four English and four orthogonal French lists for a total of 48 sentences per language. Within each group, each list was paired with one of four filler lists of the same language, consisting of a set of 26 unique non-experimental sentences. During aural presentation, the 12 experimental and 26 filler sentences in each pairing were randomly interspersed.

Experimental Blocks. The sentential factors (voice and plausibility) were crossed with two task factors, language (L1 versus L2) and load (single- versus dual-task), in a 2x2x2x2 within-subjects design. The task factors were varied by block according to eight distinct schedules that were semi-counterbalanced to control for order effects, and that were uniquely paired with one of the 8 item-list groupings described above. Each schedule consisted of four sessions held on different days, with two blocks per session. Blocks occurring within the same session differed according to both language and load. Within each schedule, participants saw four of a possible eight variations of each item frame according to the following constraints: (a) two of the four variations were in English and two were in French; (b) within each language, the repeated frame differed in terms of voice, plausibility, and load—e.g., a participant exposed to the implausible active variant in the single task condition would see the plausible passive variant in the dual task condition; and (c) no frame repetition occurred in the same language on the same day.

Procedure

Single-task blocks. Participants were tested individually using an automated paradigm with PsyScope X (Cohen, MacWhinney, Flatt, & Provost, 1993), modeled after Ferreira's (2003) task. Each trial began with the instruction *Press the black or white button to continue* displayed onscreen, followed by aural presentation of the sentence, which in turn was followed by a visually displayed prompt used to elicit sentence interpretations (see Figure 6a). For the experimental trials, ACTOR / ACTEUR elicited the agentive noun. For the filler sentences, four alternative prompts were used: (1) EVENT / ÉVÉNEMENT; (2) HOW / COMMENT; (3) WHEN / QUAND; and (4) NAME / NOM, which respectively called for a verb (e.g., *bake*), an adverb (e.g., *quickly*), temporal information (e.g., *last night*), or the proper name of the subject or object noun (e.g., *Mary*) presented in the preceding filler sentence. Participants provided an oral response following the presentation of each prompt and response times were recorded at voice onset.

During the instruction period, the experimenter defined each prompt and provided a practice example as follows: "ACTOR refers to the noun in the sentence that is doing the action. For example, in the sentence *The boy kissed the girl,* who is the ACTOR?" The session then proceeded with eight randomized practice trials with two trials per probe type. For the experimental trials, one sentence was presented in the active voice, *The butler retrieved the wine for Thomas,* and one in the passive voice, *Mars was inhabited by aliens last December.*

If participants made an error during the practice trials, the experimenter prompted them using the following sample script: "Earlier you heard the sentence *Mars was inhabited by alien*

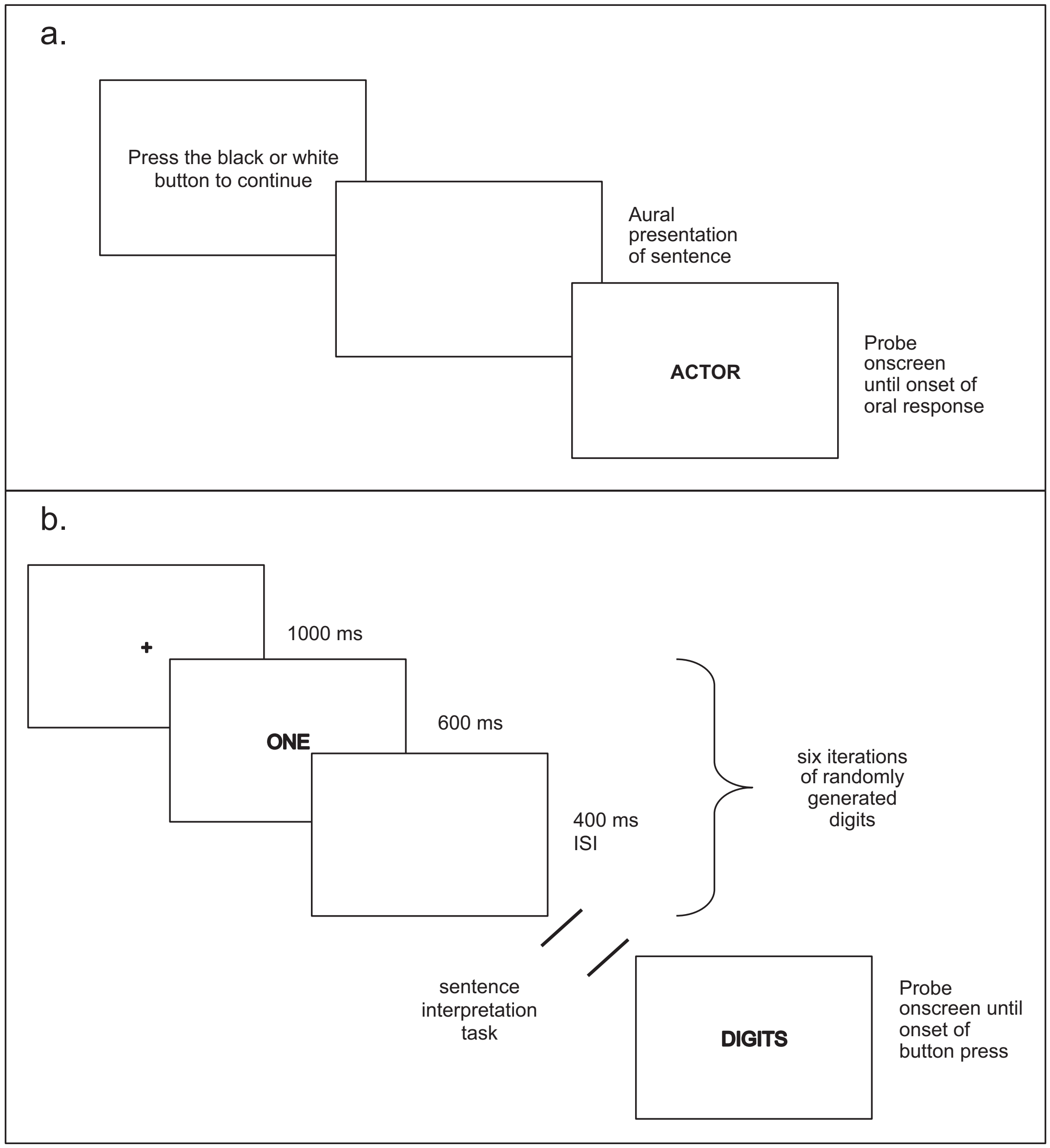

Figure 6. Trial sequences of English blocks in the (a) single-task and (b) dual-task conditions. In the dual-task condition (b), the sentence interpretation task was the same as the one depicted in (a). French blocks were identical except for the language of display.

last December. What would be considered the ACTOR in that sentence?" If the participant did not spontaneously correct their mistake upon a second prompting, their data was excluded from analysis. This occurred in 5 instances as noted in our description of participants above. This criterion ensured that participants acquired sufficient competency in both languages to conceptualize sentential agents correctly prior to the experiment. To prevent language-switching costs, instructions and practice were conducted in either English or French, as per the block condition, and each block began with three warm-up trials.

Dual-task blocks. In addition to the sentence comprehension task described above, a WM load was added to each trial as follows. After participants initiated the trial, a fixation mark was displayed for 1000 ms followed by six randomly generated digits presented visually in succession. Digits were presented lexically, e.g., THREE / TROIS, rather than numerically to control for language of activation and to prevent digit-sentence language switching. Participants were instructed to recall digits in the language of presentation. Each digit-word appeared for 600 ms followed by an inter-stimulus interval (ISI) of 400 ms. At the offset of the sixth ISI, the sentence played and comprehension was probed as in the single task condition, after which the prompt DIGIT / CHIFFRES appeared onscreen. Participants recited as many numbers as they could recall from the beginning of the trial, and then pressed a button to proceed to the next trial (see Figure 6b). The dual-task blocks thus required participants to maintain the numbers in WM throughout the comprehension task.

Results & Discussion

Data Screening

Initial screening of data revealed progressive practice effects from block to block, with Grand Mean (*GM*) accuracy increasing from 74% to 88% in a stepwise fashion from the first to the last block (see Figure 7). To our knowledge, this is the first experiment on the passive misinterpretation effect to deploy a block design with multiple sessions. Therefore, relative to other studies, our participants were exposed to a large number of sentences for each level of voice and plausibility, specifically 24 sentences per condition and 96 experimental sentences in total. In contrast, in previous experiments employing the same task and materials (Christianson et al., 2010; Ferreira, 2003; Riven et al., in preparation), each participant saw only 6 sentences per condition and 24 sentences in total. Thus, the incremental increases in performance from

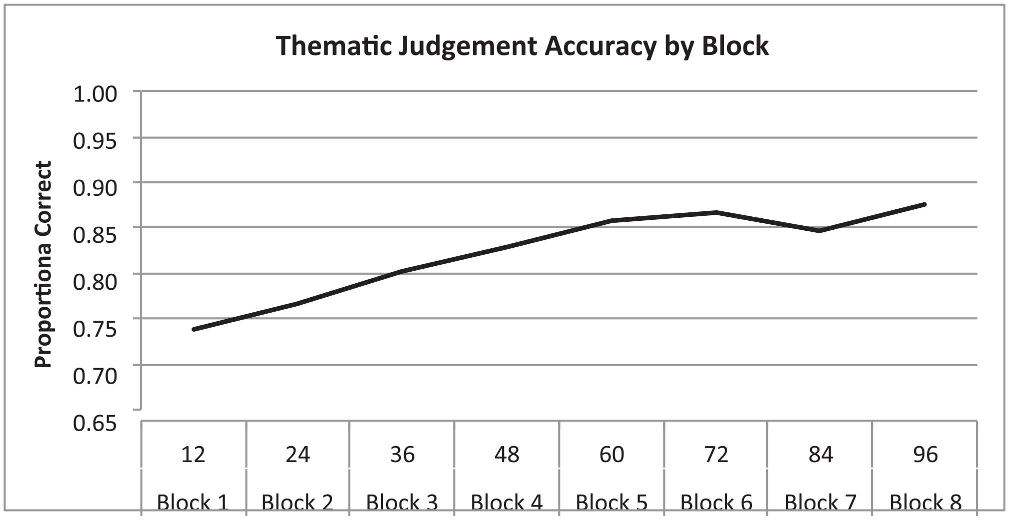

Figure 7. Grand mean accuracy in thematic judgment by block. Cumulative number of trials completed appears above the block number on the categorical axis.

block to block raises the concern that participants had developed task-specific strategies that might obscure the underlying processing mechanisms that the task is designed to assess.

The pattern in Figure 7 is consistent with effects of perceptual learning on tasks of divided attention (Shffrin & Schneider, 1977). In Shiffrin and Schneider's (1997) classic dual-task experiment, repeated exposure over a large number of trials resulted in the development of automaticity for task-specific processes, and reduced reliance on the control system for performance. Because our language-by-load manipulation was designed specifically to measure L1-L2 differences in controlled language processing, we wished to minimize the effects of task-specific learning on the experiment. We therefore excluded the last two blocks from analysis—that is, blocks 7 (*GM*=85%) and 8 (*GM*=88%), which represent an 11-point increase in performance from blocks 1 (*GM*=74%) and 2 (*GM*=77%) respectively. Appendix B includes analyses performed using all 8 blocks for comparison with the results reported below.

Trials for which participants failed to provide a task-relevant response—either the subject or object noun—were also excluded. This occurred 38 times in 3216 observations (1.18%). Removal of data did not result in significant disparities in the number of observations per condition, with n for all 16 cells falling within the range of 193 to 205. All analyses were performed using the 'lme4' package (Bates, Maechler, & Bolker, 2013) in R (R Core Development Team, 2012).

Accuracy

Response accuracy data were fitted to a mixed-effects logistic regression model with participants and items entered as random effects, and voice, plausibility, language, load, and all first-order interaction terms entered as fixed effects. The specified model provided a better fit to the data than a null model consisting of only random predictors $\chi^2(10)=244.7$, $p < 0.001$. Descriptive statistics are presented in Figure 8.

Sentential factors. Concerning voice and plausibility, we replicated the main effects found in earlier studies, with reduced response accuracy for passives compared to actives, E=4.29, SE=0.43, z=9.98, p<.001, and for implausible sentences compared to plausible sentences, E=1.28, SE=0.49, z=2.67, p=.008. The odds of correctly reporting the agentive noun for active sentences was 72.96 times that of passive sentences, 95% CI [31.42, 169.43]. The plausibility effect was more moderate, with odds of correct agentive assignment in plausible sentences being 3.59 times that of implausible sentences, 95% CI [1.41, 9.16].

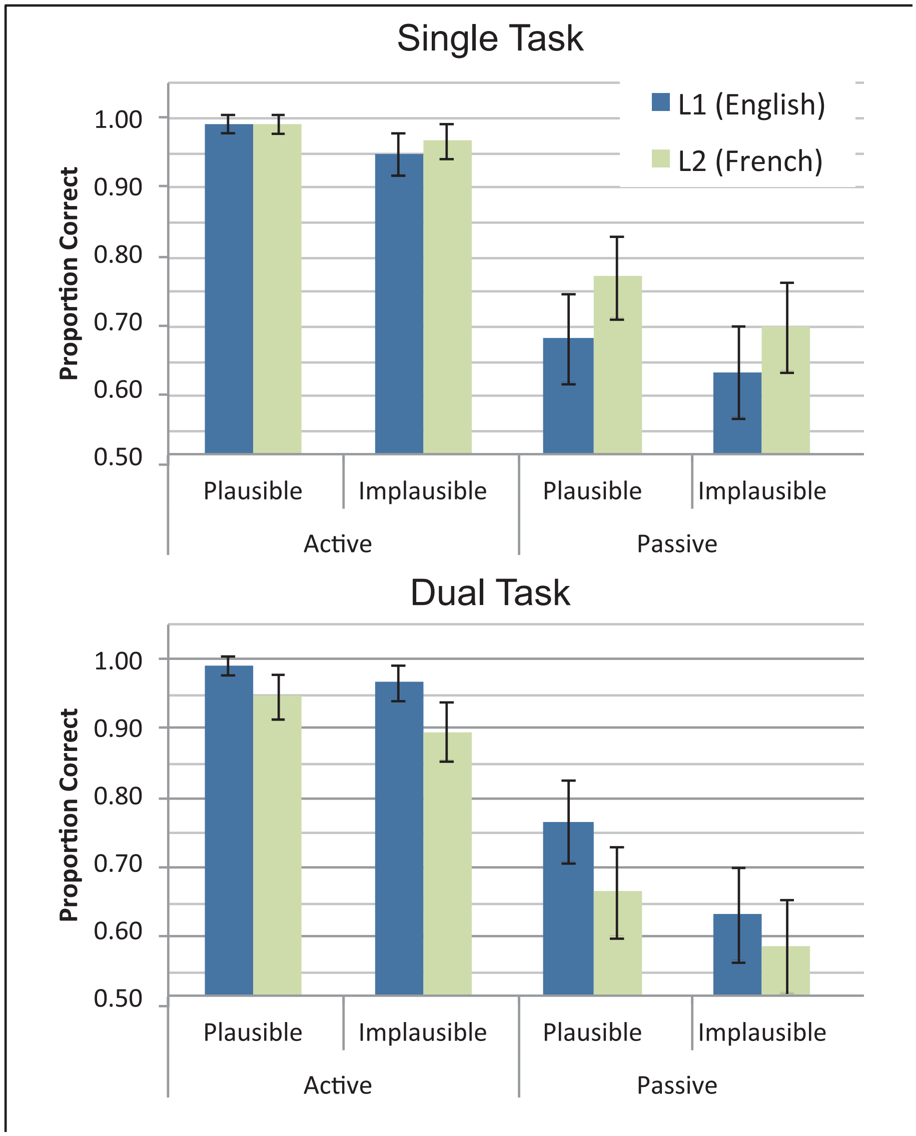

Figure 8. Proportion correct report of the agentive noun in Single Task and Dual Task conditions. Error bars represent 95% confidence intervals.

Task factors. More central to the motivations of the present experiment are the two novel task factors: language and load. The results of the fitted model supported our hypotheses. Specifically, while there was no main effect of either language, z=0.52, p=.604, or load, z=0.64, p=.524, we observed a crossover interaction between these factors, E=-0.995, SE=0.29, z=-3.43, p<.001, OR=0.37, 95% CI [0.21, 0.65]: participants were more accurate in their L2 than in their L1 when reporting the agentive noun of sentences in the single-task condition. This was especially evident for passive sentences (see Figure 8), for which the noun-position heuristic competes with the grammatical response. However, when the WM load was introduced in the dual-task condition, this pattern was reversed, owing to declines in L2 performance with relatively stable L1 accuracy. Crucially, in the dual-task condition, digit recall performance—measured in terms of percent-correct scores—was constant across levels of language, voice, and plausibility, all z's<1.19, all p's>.24, all ORs<1.16, suggesting that interference was unidirectional, from digit rehearsal to controlled language processing.

Passives only model. Given that the language-by-load manipulation is the central focus of this paper, we wished to model the interaction for passives, for which good-enough misinterpretations have been found to be most prominent in earlier studies (Ferreira, 2003, Christianson et al., 2010, Riven et al., in preparation). Table 9 presents the results of a mixed-effects logistic regression model involving only passive sentences collapsed across levels of plausibility. Accuracy was regressed on language, load, and their interaction, with random intercepts for participants and items. Comparison to a null model yields $\chi^2(3)$=8.41, p=0.038.

As in the previous model, the passive-only model reveals no main effects of load or language, but a crossover interaction between these factors (see Figure 9). Whereas L2 performance was better in the single-task condition, it was worse in the dual-task condition, owing to an 11-point drop in L2 accuracy under load while L1 performance remained relatively stable.

These results provide evidence for a decoupling of interpretation and off-line thematic assignment mechanisms. The relative resilience of L2 processing under conditions that permit deliberation suggest that task performance draws significantly on metalinguistic conceptualizations of the input—vis-à-vis which of the two nouns in the preceding sentence assumes the actor role. We argue that conceptualizing this metadata is more difficult in an L1 because in contrast to L2 processing, L1 comprehension processes are by default implicit, thus,

Table 9

Passives-only mixed-effects logistic regression model.

Fixed Effects	E	SE	z	*p*	*OR* (e^{E})	95% CI
Intercept	1.27	0.40	3.16	0.002	3.55	[1.62, 7.77]
Load	0.16	0.22	0.74	0.458	1.18	[0.76, 1.81]
Language	0.32	0.27	1.20	0.231	1.38	[0.81, 2.34]
Load x Language	-0.81	0.32	-2.50	0.013	0.45	[0.24, 0.84]

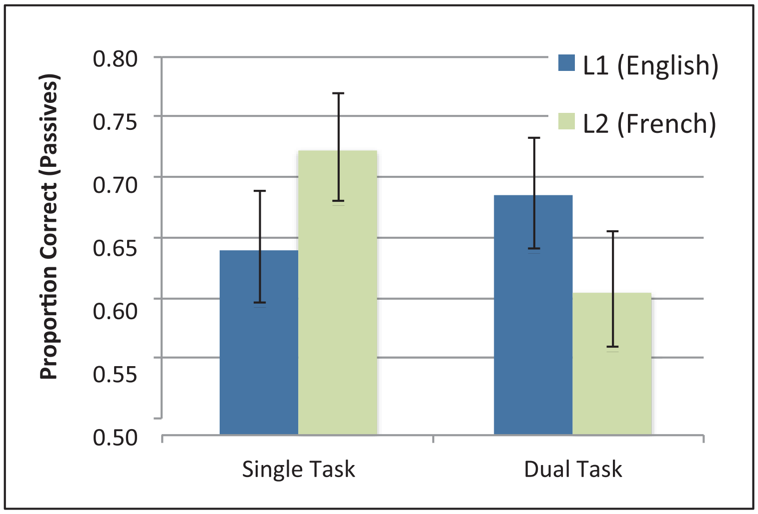

Figure 9. Mean correct report of the agentive noun of passive sentences as a function of language and load. Error bars represent 95% confidence intervals.

requiring engagement of the control system at a later stage of the task. Evidence for this hypothesis is further supported by our dual-task effects showing that when controlled processing is impeded, responses are less susceptible to interference from heuristics in an L1 than in an L2. Given that L2 comprehension is more dependent on metalinguistic processing, we suggest that, under load, a structural rendering of the sentence is relatively unavailable in an L2, thus, increasing dependency on heuristic cues for agency. Although we do not necessarily regard the 4-point bump in L1 performance under load as substantive, a possible explanation for this improvement is that digit-maintenance suppresses processing of inputs that are extraneous to L1 parsing; namely, noun-position and plausibility cues for agency.

Testing The Switch-Cost Hypothesis: Response Times and Working Memory Capacity

A central feature of our hypothesis is that L1 performance requires post-interpretive switching from linguistic to metalinguistic processes, whereas L2 performance does not due to greater dependence on metalinguistic processing for comprehension. We reasoned therefore that L2 comprehension processes may ease access to the metadata, while L1 semantic composition may engender delays associated with initiating metalinguistic analyses. If indeed L1 thematic judgments require switching from semantic compositional to metalinguistic processing, then L1 RTs should show evidence of delayed access to the agentive noun of passive sentences relative to L2 RTs.

Response Times

All RT data and analyses were restricted to correct trials and data were normalized via log transformation. Observations falling beyond +/- 2.5 SDs of mean logged RTs were removed (n=51, 1.97%), as were an additional 16 observations falling below a raw RT of 500 ms (0.63%). Because RTs were recorded at voice onset, it is likely that responses below 500 ms reflected non-specific vocalizations unrelated to the response. Figure 10 shows mean RTs and 95% CIs in the original metric for ease of interpretation.

We regressed logged RTs on voice, plausibility, language, task, and all first order interaction terms in a mixed-effects linear regression model, with items and participants entered as random effects. Comparison to a null model consisting of only random effects showed that the full model provided a better fit to the data $\chi^2(10)=249.7$, $p<0.001$. Table 10 summarizes the statistically significant estimates of the model. All other effects were non-significant (all *p*'s > .09).

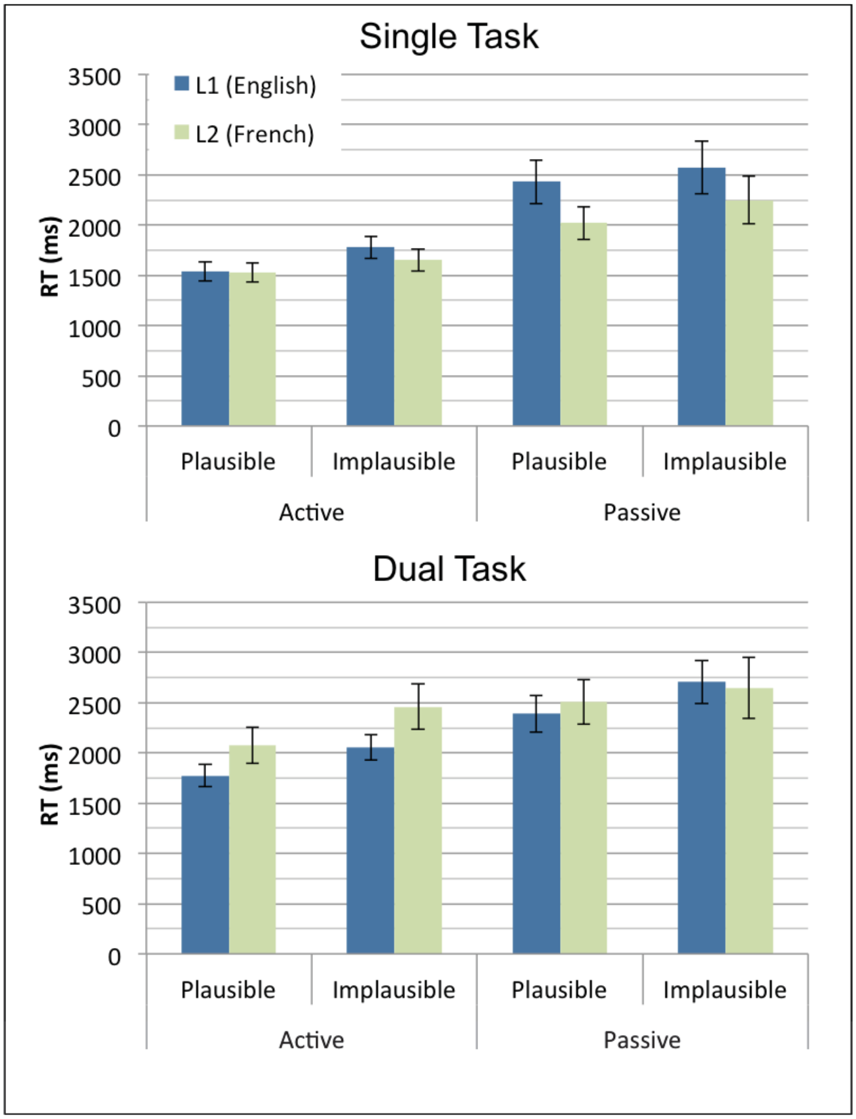

Figure 10. Mean RTs in Single Task and Dual Task blocks as a function of voice, plausibility, and language. Error bars represent 95% confidence intervals.

Table 10

Mixed-effects linear model of logged response times.

Fixed Effects	E	SE	t	*p*	95% CI
Voice	0.306	0.039	7.82	<0.001	[0.23, 0.38]
Plausibility	0.119	0.037	-3.23	<0.001	[0.19, 0.05]
Load	0.204	0.032	6.30	<0.001	[0.14, 0.27]
Voice x Load	-0.106	0.034	-3.17	<0.001	[-0.17, -0.04]
Load x Language	0.091	0.035	2.57	0.005	[0.02, 0.16]
Voice x Language	-0.123	0.040	-3.06	0.001	[-0.2, -0.04]

Results show significantly increased RTs for passives compared to actives (voice), for implausible compared to plausible sentences (plausibility), and for dual task compared to single task conditions (load). In addition, we found a voice-by-load interaction, whereby the effect of voice (increased RTs for passives compared to actives) is less pronounced in the dual task condition. This is likely because dual task RTs also reflect digit processing, which may wash out language specific effects, for instance by elevating RTs for all sentence conditions. As in the accuracy data, we also found a language-by-load interaction, suggesting that the delays associated with the dual task condition are more pronounced in an L2. Crucially, however, this finding was largely driven by the fact that L2 processing was faster in the single task condition (see Figure 10), but increased more sharply in the dual task condition. This RT effect largely mirrors what was observed for the accuracy data: bilinguals performed better in their L2 when the control system was unencumbered, but worse in the presence of interference from non-linguistic sources. That bilinguals were faster in their L2 in the single task condition is particularly relevant to our switch-cost hypothesis, because L1 semantic compositional processes should be faster in an L1 than an L2. The fact that L1 RTs were slower in the single task condition, suggests that processing delays are likely associated with the difficulty of generating a metalinguistic judgment, rather than interpretation. In support of this hypothesis, we found a voice-by-language interaction showing that the delays associated with passive processing were more pronounced in the L1 condition, and this was largely driven by differences in single task performance (see Figure 10). This pattern is consistent with the hypothesis that correct report of the agentive noun entails accessing metalinguistic representations post interpretation. We propose that in an L2, these metalinguistic representations are constitutive of sentence processing routines executed by the control system, whereas in an L1 these metalinguistic representations are extraneous to the default parsing mechanism. Therefore, whereas L2 processing facilitates access to the metadata, L1 processing engenders switch costs when controlled analysis of the input is required.

Working Memory Capacity

Additional evidence for the L1 switch-cost hypothesis was obtained from participants' working memory capacity (WMC) scores, which were assessed in an independent reading-span task (Conway et al., 2005) conducted after the experiment on the final day of testing. Following the same rationale as above, WMC should be associated with performance in both languages given the controlled nature of the agent-assignment task, but possibly to a greater extent in L1 due to additional switching demands. Using mixed-effects models, we found support for our hypothesis in the form of significant associations between WMC and (a) global response accuracy, E=0.02, SE=0.004, z=4.60, OR=1.02, 95% CI [1.01, 1.03] and (b) single-task RTs, E=-0.004, SE=0.001, t=-3.75, 95% CI [-0.01, -0.001]. In addition, we observed a WMC-by-language interaction for single-task RTs, with the association between these factors being more pronounced in L1 than L2, E=-0.001, SE=0.001, t=-2.06, 95% CI [-0.003, -0.00001].

Although these mixed-effects model estimates were statistically significant, we note that they were smaller than we would have expected, which may restrict their generalizability. Simple Pearson correlations of WMC and accuracy scores, averaged by participant (N=45), in all language-by-load conditions—i.e., L1-single-task, L1-dual-task, L2-single-task, L2-dual-task—were not statistically significant, max r=.22, min p=.15, nor were the correlations of WMC with logged RTs in the same conditions, max r=|-.19|, min p=.21. Nevertheless, given prior evidence that WMC is associated with controlled language performance (e.g., Just & Carpenter, 1992), and given the face validity for such an association, we suggest that the statistically significant, though marginal, effects reported above reflect true relations. However, the magnitudes of these associations in our sample may have been minimized due to measurement error of WMC. Here it is worth noting that participants completed the reading span task on the final day of testing following an experimental session that lasted roughly 45 minutes. It is therefore possible that increased fatigue or reduced motivation to engage in yet another demanding task rendered WMC scores less reliable than they would have been under more ideal circumstances.

Discussion

A growing number of studies have shown that native speakers do not always generate the correct interpretation of sentences on comprehension tasks (Christianson et al., 2001; Ferreira, 2003; Frazier & Clifton, 2015), suggesting that semantic compositional processes rely on extra-grammatical heuristics to determine how the input maps onto its structure (Ferreira et al., 2002; Ferreira & Patson, 2007; Karimi & Ferreira, 2015). Evidence that native speakers routinely misconstrue the agents of grammatical and unambiguous passive sentences (Ferreira, 2003) has further contributed to a view of syntactic processing as effortful, slow, and vulnerable to interference (Karimi & Ferreira, 2015). We suggest however that while there are certainly various aspects of language processing that require effortful and deliberate analysis, the parsing mechanism is automatic and in turn semantic composition is generally felicitous and correct. With respect to the passive misinterpretation effect, we propose that errors reflect response interference when participants attempt to generate a post hoc metalinguistic formulation of the sentence's thematic roles.

Our results provided evidence that indeed off-line thematic assignment errors reflect metalinguistic processing post interpretation. Specifically, we showed that differences in the underlying representations of first and second languages (Paradis, 2009; Ullman, 2001) modulate the effects of heuristic cues on task performance. When comprehension is dependent on metalinguistic analysis—as is the case in an L2—bilinguals show enhanced control of competing heuristic and structural cues for agency, but succumb increasingly to heuristics when controlled processing is impeded. In contrast, when comprehension is dependent on implicit semantic compositional processes—as is the case in an L1— bilinguals' conceptualization of thematic roles is more susceptible to heuristic cues for agency, owing to switch costs associated with activating metalinguistic representations at a later stage of analysis. However, L1 comprehension shows no further decrements as a function of interference from non-linguistic sources.

Our results support the presence of a privileged layer of semantic compositional processing in a native language that generates meaning independent of the control system, above which resides discrete metalinguistic knowledge that is recruited specifically when deliberate analysis is required. Consider for instance that when editing a manuscript, one may experience the intuition that a sentence contains an error without conscious access to the rule or norm that is being violated. To improve the sentence, one launches a secondary process of conceptualizing and correcting the error. This may involve applying grammatical agreement rules that one has explicitly memorized, or trial-and-error testing of edited clauses against one's "ear". Either way, the knowledge that first generates the intuition is not directly accessible to the metalinguistic analysis that ensues. Our data suggest that good-enough comprehension is governed by a similar duality in language processing—namely a division between semantic-compositional processes and metalinguistic thinking—and that it is strictly the latter form of analysis that is effortful and slow.

We conclude that the extra-grammatical heuristics, which have been shown to bias interpretations, provide a fast-and-frugal supplement to metalinguistic analysis—not interpretation—and may be confined to outcomes that require thinking about the input. We support the view that the language comprehension system might yield a “good-enough” interpretation in some instances, but as we have shown, the heuristics that drive the passive misinterpretation effect do not necessarily impinge on the underlying semantic-compositional process, but instead derail metalinguistic judgments, which corrupt retroactively the outputs of the language comprehension system.

Chapter 5: General Discussion

Several experiments have shown that native speakers sometimes report incorrect interpretations of sentences (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira, 2003; Frazier & Clifton, 2015), leading to the proposal that the language comprehension system sometimes relies on error-prone heuristics, in favour of structure, to determine the meaning of a sentence (Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007). According to the good-enough approach, the processes required to fully integrate the structural elements of a sentence are too difficult to compute in some circumstances, resulting in a preference for "fast and frugal" heuristics which may provide a sufficiently satisfactory—that is, a "good enough"—representation of the input (Karimi & Ferreira, 2015). The apparent misinterpretation of passive sentences on thematic assignment tasks (Ferreira, 2003) has served as one of the chief lines of evidence that extra-grammatical heuristics may supersede structure in the determination of sentence meaning. The present thesis questioned this evidence by asking: Do these heuristics supersede structure in the course of semantic composition, or do they interfere with metalinguistic judgments about the properties of a sentence?

This line of inquiry was motivated in part by scepticism that structure is more difficult to process than extra-grammatical determinants of comprehension. In their description of the language comprehension system, Ferreira and Patson (2007) draw analogies to judgment and decision-making—a cognitive domain that is notoriously prone to error due to the difficulty associated with probabilistic reasoning and a reliance on heuristics when formulating probability judgments (Tversky & Kahneman, 1974). Much like dual-process models of decision-making (Evans & Stanovitch, 2013), the good-enough approach proposed that sentence interpretations are governed by a division of labour between heuristic and algorithmic information processes. While the heuristic stream computes meaning from extra-grammatical cues in the input, the algorithmic stream is responsible for processing its structure (Karimi & Ferreira, 2015). However, the analogy to judgment and decision-making breaks down when we consider that what is effortful and slow about probabilistic algorithms is that they require explicit knowledge and therefore load on limited working-memory (WM) resources (Evans & Stanovitch, 2013). But this characterization of algorithmic processing presumably does not apply to how syntax is processed in a native language. Native speakers are rarely required to consciously think about how grammar (or other formalized linguistic classes, such as thematic roles) relate to the meaning of a sentence. Grammatical knowledge is implicit and is generally integrated without the need for deliberate processing, except perhaps for circumstances in which there is undue complexity or ambiguity in the signal. But passive sentences—excluding for the moment implausible ones—are grammatical, unambiguous, and relatively common (Roland, Dick, & Elman, 2007), raising the question of whether the structural computation of a passive sentence is indeed as effortful as the good-enough approach implies. Could it be that the passive misinterpretation effect (Ferreira, 2003) does not reflect comprehension errors, but rather a lapse in judgment about the formal properties of a sentence? It seemed possible that the good-enough approach was conflating two different notions of structural processing: (1) application of implicit phrase structure rules in the course of semantic composition—a largely automatic process—with (2) retrieval of explicitly memorized prescriptive, or metalinguistic, rules—a controlled process. Because off-line comprehension tasks such as Ferreira's (2003) measure metalinguistic thinking, it was hypothesized that the passive misinterpretation effect may arise in the course of metalinguistic processing, not semantic composition.

The methodological strategy for testing this hypothesis was to investigate whether the passive misinterpretation effect could be modulated depending on whether sentences are presented in the interpreter's first (L1) or second language (L2). Because L2 comprehension is more dependent on controlled metalinguistic processing and L1 comprehension is more dependent on implicit linguistic processing (Paradis, 2004; 2009; Ullman, 2001; 2015), L1 and L2 comprehenders may show different degrees of dependency on heuristics in comprehension tasks. A WM load manipulation was also applied to determine the degree to which L1 and L2 performance may be further disrupted by non-linguistic task demands. As detailed in the preceding chapters, these manipulations provided evidence that the passive misinterpretation effect is modulated by controlled processing dynamics (Manuscript 1, Chapter 2), and that L1 comprehension of passives may ensue felicitously, without contributions from the generalized control system (Manuscript 2, Chapter 4). As will be outlined below, the results suggest that language-processing heuristics are not necessarily easier to compute than structure and may impinge on performance primarily for metalinguistic aspects of language processing. In the next section, I briefly summarize the results of the experiments in Manuscript 1 and Manuscript 2 before elaborating on the implication of these results for language processing more generally.

Summary of Results and Interpretations

Manuscript 1. The results of two between-subject experiments showed that success on Ferreira's (2003) thematic assignment task draws on the generalized control system. In a single task experiment, there was no difference between L1 and L2 groups in reporting the agents of sentences. Both groups showed the same pattern of successes (and failures) as a function of voice and plausibility—with increased errors for passives and implausible sentences. In contrast, in a dual-task experiment (comprised of a new group of participants that did not participate in Experiment 1), language group modulated this pattern, with the L2 group reporting fewer errors than the L1 group as a function of voice and plausibility. Although the direction of the effect was surprising, it occurred because the L1 group appeared to have allocated greater attention and effort to the WM component of the dual task experiment (i.e., digit recall), performing better than the L2 group by approximately 10 percent points across all levels of voice and plausibility (see Figure 5, Chapter 2, p.35). These trade-offs suggested that success in reporting the agents of passive sentences was dependent on generalized controlled resources.

Although we speculated that these controlled resources are likely involved in metalinguistic processing rather than interpretation, evidence for this conclusion was not directly obtained due to task-selection effects, which obscured the underlying mechanism of L1-L2 differences. For instance, the results may also be consistent with a view that would take L1 and L2 comprehension processes to be equally dependent on the control system. Although this interpretation contradicts what is known about the attentional demands of L2 processing (Clahsen & Felser, 2006; Segalowitz, 2010), it is nevertheless prudent to restrict interpretations to what had been reliably demonstrated by the data: the passive misinterpretation effect varies as a function of controlled processing effort. Thus, although the good-enough approach does not explicitly link the algorithmic stream to the control system, Manuscript 1 showed that the processes required for generating a veridical response indeed engaged the control system, opening up the possibility of a division between implicit semantic compositional processes and explicit thinking in good-enough performance.

Manuscript 2. To minimize the task-selection effects found in Manuscript 1 and eliminate potential between-subjects confounds, the research reported in Manuscript 2 used a within-subjects design with English-French bilinguals. Each participant performed the agent identification task in their L1 and L2 in both single and dual task conditions in one experiment.

The results showed a crossover interaction between language and load, with better L2 performance in the single task condition, but better L1 performance in the dual task condition. Crucially, digit recall on the WM load task was stable across languages. With respect to the single task results, we attributed the edge in L2 performance to a *metalinguistic buffer effect*, whereby thematic judgments were enhanced due to higher engagement of metalinguistic representations during interpretation. This hypothesis was further supported by reaction time (RT) and working memory capacity (WMC) data, which respectively showed that L1 agentive judgments for passives were slower and more strongly correlated with WMC than L2 agentive assignments in the single task condition. Here, the within-subjects design is worth highlighting. The L2 edge in both speed and accuracy, as well as the reduced reliance on WMC, occurred in the same individuals with the same cognitive abilities. This strongly suggests that L1-L2 differences occurred primarily as a function of how first and second languages are processed. We proposed that L2 thematic judgments are facilitated by explicit metalinguistic processing. In contrast, L1 thematic judgments may be forestalled due to initial reliance on implicit semantic composition strategies and the need to initiate metalinguistic analysis at a later stage of processing.

With respect to the dual task results, the crossover interaction provided additional support for the hypothesis that L1 comprehension processes are not particularly taxing by showing that there were no added decrements in L1 performance associated with a WM load. In fact, L1 accuracy was numerically improved under load, while L2 performance declined precipitously. This illustrated that the *metalinguistic buffer effect* initially observed in the single task condition could be eliminated by impeding the control system, and that the semantic compositional processes which contribute to correct agentive judgments in a native language operate independently of the control system. Collectively, these results provided evidence for two distinct levels of processing that contribute to thematic judgments; namely, (a) semantic compositional processes involved in computing an interpretation of the input, and (b) controlled processes involved in formulating a metalinguistic judgment of the output. While these data replicated Ferreira's (2003) passive misinterpretation effect in all language-by-load conditions, the evidence suggested an alternative explanation for when and how heuristics come to supersede veridical processing. That is, in a native language, heuristics do not supersede structure in the determination of sentence meaning, but supplant controlled metalinguistic processes that participate in the formulation of thematic judgments.

Controlled Mechanisms of Thematic Judgments in Actives and Passives

Although I argue that passive sentences are interpreted easily and correctly, the passive misinterpretation effect is a robust phenomenon, having been replicated in five different samples in the current research program, and in at least two others from previous studies (Christianson et al., 2010; Debrawska & Street, 2006). The effect may be described in theory-neutral terms as a bias towards reporting the first noun in a sentence as the agent. But if it is the case that this bias is decoupled from the interpretation of the sentence, then the effect begs for an alternative to the good-enough explanation. Why should agentive judgments consistently drift towards the first noun of a passive sentence if interpretations are correct? Although linearity and plausibility cues are certainly at play, one might expect correct interpretations to supersede these cues if the compositional meaning of a sentence is in fact obtained. It is therefore important to propose a possible mechanism through which controlled thought processes may be systematically biased, even when interpretations are not. Here, I will propose two such mechanisms—(1) *topic anchoring* and (2) *predict and revise*—each of which may contribute some measure of variance to the error-rates consistently observed for passives on agent judgment tasks.

Topic anchoring. Active and passive sentences differ not only with respect to their structure, but also with respect to pragmatics, specifically concerning the mapping of thematic roles to the *topic* versus the *comment* of a sentence (Gundel & Fretheim, 2004). In English, the *topic*—the main referent of the discourse—is usually placed at the beginning of a sentence, and *comment*—the predicate, or that which is being said about the topic—is usually placed at the end. Active and passive sentences thus differ in terms of what is treated as the main, or important, referent of the event (Johnson-Laird, 1968). Because speakers place what they are talking about at the beginning of a sentence, actives are construed as being *about* their agents, while passives are construed as being *about* their patients. Thus, in speech production tasks, the use of the passive voice is more likely when the patient noun, i.e., the topic, is given as a production cue (Bock, 1986). That agents may be altogether omitted from passives—as in the sentence *The employee was fired*—further illustrates that passives are used to cast the patient as the primary referent of the event.

One possible consequence of topicalizaiton is that even when agents are explicitly mentioned in passive sentences—as in *The employee was fired by the manager*—the patient noun waxes in the attentional spotlight, while the agent wanes in the periphery. This may have systematic effects on how interpreters think about the information given by a sentence. For instance, when asked to colour in a rectangle to match such sentences as *Red precedes blue* (active) and *Blue is preceded by red* (passive), participants allot more space to the colour that is topicalized by the sentence, resulting for instance in greater allotment of *red* to the active and of *blue* to the passive (Johnson-Laird, 1968). It has also been shown that a *topic* noun provides a better retrieval cue than a *comment* noun on a delayed sentence recall task (Perfetti & Goldman, 1975). These kinds of effects suggest that the conceptual representation of a sentence may be anchored to its topic, which may provide the primary access point for non-linguistic processes. With respect to the agent identification task, the correct response may be more difficult to retrieve for passives than for actives because it requires inhibiting the more salient constituent, and this executive difficulty should persist even when interpretations are correct.

Although this thesis was not designed to test the *topic-anchoring* mechanism, data from previous studies (Christianson et al., 2010; Ferreira, 2003) provide some indirect support, specifically from performance on patient (compared to agent) identification questions. Both Ferreira (2003) and Christianson et al. (2010) measured performance for patient probes, the results of which are presented in Table 11 (which also includes a summary of performance on agent probes, including those from the present thesis). As can be seen, the passive misinterpretation effect is less pronounced when participants are cued to report the *patient* of a sentence (compare grand mean accuracy for agent [top] versus patient [bottom] judgments). The pattern shows moreover an asymmetry between passive and active sentences for agent and patient judgments: for passives, patients are easier to retrieve than agents, but for actives, agents are easier than patients. For instance, Christianson et al. (2010) found that while participants performed at ceiling when reporting the agents of plausible actives such as *The dog bit the man,* they misconstrued the patients of these sentences 22% of the time. (The same pattern is observed for implausible actives.) This result is difficult to reconcile with a view that would take thematic judgments to be a direct index of comprehension, because this would entail that simple transitive actives are systematically misinterpreted, even when linearity, plausibility, and structural cues in the input all converge on the same construal of the event—i.e., that a dog bit a man. It seems

Table 11

Thematic judgment data for the same materials from four independent studies.

	Active		Passive	
	Plausible	Implausible	Plausible	Implausible
Agent Judgments (GM[a])	**0.99**	**0.96**	**0.84**	**0.72**
Ferreira (2003)	0.99	0.99	0.88	0.74
Christianson et al. (2010)	0.99	0.91	0.90	0.74
Manuscript 1 (L1-single)	0.99	0.98	0.89	0.78
Manuscript 2 (L1-single)	0.99	0.95	0.67	0.62
Patient Judgments (GM[a])	**0.88**	**0.84**	**0.88**	**0.80**
Ferreira (2003)	0.97	0.91	0.92	0.85
Christianson et al. (2010)	0.78	0.77	0.83	0.75

[a] GM=grand mean accuracy, averaged across studies.

Note. Data presented from Manuscript 1 and Manuscript 2 of this thesis are restricted to first-language and single-task conditions (L1-Single) to maintain comparability with Ferreira (2003) and Christianson et al., (2010).

rather that the difficulty of retrieving the patient of an active sentence—and the agent of a passive sentence—arises when the required role is not the main referent of the event. The asymmetry between agent and patient judgments for passives and actives suggest that regardless of structure or thematic hierarchy, the topic is easier to access than the comment, and thus, some proportion of variance is associated with pragmatic factors that impinge on judgments beyond the realm of comprehension.

Predict and revise. Another possible mechanism by which passive sentences might engender executive control difficulties despite correct interpretations is introduced by the dynamics of incremental processing. Sentences unfold over time, but the language comprehension system does not passively wait for a sentence to end before computing an interpretation. Garden path effects are exemplary of incremental dynamics (Frazier, 1987). The difficulty of integrating the second clause in a sentence like *While Anna dressed the baby played in the crib* arises because the parser has already assigned its subject—i.e., *the baby*—to the preceding clause. This illustrates that one's understanding of the input is a function of the information available at the moment of computation, which may be dynamically updated as the signal unfolds—a mechanism which I will refer to as *predict and revise*.

According to the *predict and revise* proposal*,* incremental processing may generate different thematic outputs depending on when the computations are performed with reference to the input stream. Specifically, thematic roles computed over a subset of the input—e.g., *The dog*—based on what is most likely at the time—i.e., that the *dog* is the agent of the incoming sentence—may be later revised when the ensuing input introduces incompatible information—e.g., *was bitten by the man.* At this point in processing*,* the language comprehension system may reassign the agentive role to *the man,* yielding the syntactically licensed interpretation*. Predict and revise* may therefore generate conceptual duality, with the output containing both a canonical interpretation rendered predictively as well as the correct agent rendered upon revision. Even if the structure is felicitously and faithfully updated by the time the full input is processed, the predicted agent may linger in the conceptual output, generating interference (akin to the dynamics of garden-path sentences; see Christianson et al., 2001; Slattery, Sturt, Christianson, Yoshida, & Ferreira, 2013). The duality that results from *predict and revise* creates a role for executive control processes because the agent rendered predictively must be inhibited to retrieve the agent rendered faithfully when generating a metalinguistic formulation of the agentive noun.

As was the case with the *topic anchoring* mechanism, we did not specifically test the *predict and revise* proposal, but the results in Table 11 may also provide some indirect support, specifically by illustrating that thematic judgments are sensitive to incremental processing dynamics. Here, I draw attention to the asymmetry between the effects of voice and plausibility on thematic judgments. Looking at the agent judgment data, it appears that implausibility has relatively little impact on active structures across samples (all means ≥ 91%, *GM*=96%). Implausibility primarily skewed outcomes for passive sentences, suggesting that semantic cues for agency take effect only once a bias towards the opening noun has already been established. I do not suggest that this is because semantics have no top-down influence on structure *tout court*. That issue is beyond the scope of the present discussion. Instead, I offer a relatively banal explanation for the pattern: plausibility cues for agency only become fully available at the end of the sentence when both nouns can be considered in relation to the verb. Compare, for instance, (1a) to (1b), whereby what we may consider implausible is not simply that *the dog was bitten*, but that it was *the man* who did the biting. Similarly for the actives in (1c) and (1d), the implausibility of a man biting something becomes apparent only when the object is revealed at the end of the sentence. Left-to-right incrementally constraints, thus, imply that plausibility effects must be at least partially suspended until the end of the sentence, after a considerable amount of processing has already taken route.

(1) a. The dog was bitten by the wolf.
b. The dog was bitten by the man.
c. The man bit his tongue.
d. The man bit the dog.

In contrast, linearity cues for agency as well as *topic anchoring* effects on attention may be forged relatively early in the sentence—in principle as early as the first NP—and might therefore skew judgments well before the entire scope of the input is processed. Linearity cues thus have a clear field advantage over semantics when processing dynamics are considered incrementally, resulting in more robust effects of voice than of plausibility. That is, while implausibility has marginal impact on agentive judgments for active sentences (*GM* error-rate = 4%), passive voice has considerable impact on agentive judgments for plausible sentences (*GM* error rate = 16%). The asymmetry suggests that thematic judgments obtained downstream are sensitive to the dynamics of incremental processing. The picture that emerges when all sentence conditions are considered relative to one another is that thematic judgments not only reflect the presence of heuristic cues in the input, but also the dynamics of when those cues become available in real time, and the potential primacy effects such cues impose on executive processes downstream.

The *predict and revise* mechanism is consistent with a variety of approaches to parsing, including garden path (Fodor & Innoe, 2000; Frazier & Ryner, 1982; Slattery et al., 2013), prediction (Altmann & Markovic, 2009; Ferreira & Lowder, 2016; Tenenhaus et al., 1989), and Bayesian models (Hale, 2011; Levy, 2008). Although these approaches vary considerably with respect to a range of theoretical issues, they all converge on the principle that commitments made early in the sentence create structural or conceptual dependencies that influence processing later on. The notion that thematic roles in particular are rendered segment-by-segment as a sentence unfolds is consistent with the extended argument dependency model (eADM; Bornkessel & Schlesewsky, 2006), which postulates a post-structural stage of processing in which (a) initial assignments are repaired and (b) diverse information sources are integrated. For the present purposes, the most relevant implication of incremental dynamics is that even if a veridical representation is achieved by the time all of the elements of the input are integrated, the initial computations may linger in the output (à la Slattery et al., 2013), introducing undue inhibitory demands on the control system when thematic judgments are cued at sentence offset.

The main distinction between the good-enough account and the two mechanisms proposed here—*topic anchoring* and *predict and revise*—is that only the latter mechanisms allow for the interpreter to have achieved and maintained a fully specified understanding of the input and still experience difficulties rendering a thematic judgment. This is advantageous because participants in fact report the correct interpretation most of the time—an observation that is difficult to reconcile with accounts that describe syntactic representations as somehow more difficult to compute or maintain than non-syntactic representations. But perhaps more importantly, there is new evidence showing that participants' erroneous reports of thematic roles may be dissociated from their otherwise correct interpretations of passive sentences. Bader and Meng (2016) recently presented data to this effect. Specifically, they showed that in a speeded plausibility judgment task, participants who had correctly categorized implausible passives—i.e., by correctly rejecting them as implausible in the vast majority of instances (i.e., 89%)—nevertheless erred 29% of the time when subsequently reporting their agent and patient nouns. The divergence between these two measures suggests firstly that a faithful rendering of passive sentences may not be especially difficult to generate and secondly that the failure to report thematic roles downstream may be decoupled from the participants' ability to fully process passive structures.

Implications for Language Processing and the Good-Enough Approach

Although the passive misinterpretation effect (Ferreira, 2003) has served as one of the primary lines of evidence that "fast and frugal" heuristics sometimes supersede structure in the determination of sentence meaning, it is certainly not the only blunder to provide evidence for the good-enough approach to language processing (Ferreira et al., 2002; Ferreira & Patson, 2007; Karimi & Ferreira, 2015). In Chapter 1 of this thesis, I reviewed a variety of such phenomena, namely, the Moses Illusion (Erickson & Mattson, 1981), lingering misinterpretations of garden path sentences (Christianson et al., 2001), and failures to notice dysfluencies in syntactically blended utterances (Frazier & Clifton, 2015). In each case, the judgment reported by the participant contradicted the compositional meaning of the input. While the present thesis provided evidence that one phenomenon—thematic misjudgment of passives—may not involve misinterpretations *per se*, this does not necessarily challenge the claim more generally that the language comprehension system occasionally outputs errors. Surely, there are circumstances in which the input structure cannot be fully integrated by the parser, as might be the case for such highly convoluted sentences like *The teacher the boy the girl hugged tripped fell*. What this thesis aims to highlight more specifically with respect to the good enough model is that its characterization of structure—as algorithmic, slow, computationally demanding, and vulnerable to interference—is too simplistic, and that not all language processing failures can be attributed to a strict dichotomy between syntax and heuristics.

While some aspects of sentence processing (including processing of structure) may be difficult in some circumstances, the present thesis underscores the importance of operationalizing difficulty not simply in terms of structural versus non-structural processes, but also in terms of automatic versus controlled processes. In contrast to the good-enough model, I propose that the linguistic system operates autonomously and yields fully compositional interpretations as long as the input signal conforms to the grammar in a clear, simple, and unambiguous manner. However, a variety of inputs may be so complex, or divergent from the grammar, that they trigger systematic “misparses”. Garden-paths provide a classic example of incremental parsing gone awry, whereby the input triggers an attachment that turns out to be incorrect. Although the attachment may be repaired eventually, the interference generated by the initial parse illustrates that the linguistic system deploys rapid syntactic strategies that sometimes result in errors. I propose moreover that when such parsing errors occur, the interpreter may launch an explicit analysis of the input that aims to diagnose and remedy the problem. These two stages of processing involve two different systems of analysis: the first system, i.e., the linguistic system, autonomously computes structure based on implicitly stored perceptual strategies, and the second system, i.e., the metalinguistic system, accesses explicit knowledge to guide deliberate interpretation. Thus, while the parser may occasionally output a non-veridical interpretation, the computationally demanding aspect of performance arises only in the second stage of processing, when the metalinguistic buffer aims to integrate parsing outputs with a global interpretation of the sentence.

What differentiates this view from the good-enough model is firstly that the initial attachment and the reparation may involve linguistic and non-linguistic systems of analysis in separate stages of processing, and secondly that it is the latter system that is slow and unstable. While native speakers rely primarily on the linguistic system to automatically output correct interpretations, the metalinguistic system may be called upon when the input signal is complex or ambiguous. In such circumstances, successful interpretation may be arbitrarily determined by what the interpreter explicitly knows or believes about the rules and norms of language use, as well as his or her motivation to carry this form of analysis through to completion. Crucially, as argued extensively above, this does not entail that all language-processing errors are interpretation errors. In the case of implausible passives for instance, correct interpretations may be autonomously computed by the linguistic system, and errors may specifically occur as a function of how the outputs are used by the buffer for non-linguistic thought processes.

Implications for Second Language Processing

The methodology for this thesis was informed by a theory of second-language processing that postulates differences in how first and second languages are stored and used. In particular, Ullman (2001) and Paradis (2004) have both described a Declarative-Procedural model of language processing, according to which the linguistic system coopts these neurological memory systems for different components of language processing. The proposal is that in a native language the procedural system stores and coordinates syntactic rules, while the declarative system stores and coordinates lexical items and their meanings. This organization, however, is different for a second language. According to the model, because learning a second language requires explicit memorization of grammatical rules, L2 representations—both the lexical-semantic and the morpho-syntactic—are stored and coordinated by the declarative memory system. It is this factor that creates L2 dependency on the control system for comprehension, because retrieving and using declarative (or metalinguistic) knowledge entails deliberate processing. While it is beyond the scope of this thesis to comment on many aspects of this theory, the results of Manuscript 2 appear to support at least one notion: grammatical knowledge is relatively more explicit in an L2 than an L1, and such explicit knowledge is coordinated by a system that is different from L1 parsing.

As noted in Chapter 1, several studies have reported instances in which linguistic judgments were found to be superior among non-native speakers compared to native speakers (Chipere, 1998; Debrawska & Street, 2006; Sasaki, 1998), suggesting that explicit L2 representations may enhance certain aspects of language processing. To my knowledge, the current research is the first to test the metalinguistic mechanism prospectively, and Manuscript 2 is the only one to find such an effect using a within-subjects design. These data thus provide unique behavioural evidence that indeed L2 grammatical knowledge is ore explicit, and dependent on a system of analysis that is dissociated from the native linguistic system. This is consistent with the aspect of the declarative-procedural model, which maps metalinguistic knowledge, be it in an L1 or an L2, to declarative memory. Crucially, the present thesis does not speak to the role of procedural memory in native language processing, and in fact, the data may be relatively more consistent with a modular view of parsing (Fodor, 1983), for which the grammar is stored and coordinated by a domain-specific linguistic system, rather than the more generalized procedural memory system.

As a final note, the *metalinguistic buffer effect* does not likely speak to L1-L2 differences at all levels of proficiency, but rather captures relative differences that may fluctuate with variation in age of acquisition and regularity of usage. I propose that the likelihood of observing a *metalinguistic buffer effect* in bilinguals may be an “inverted-U” function of L2 fluency, whereby the accessibility of explicit grammatical knowledge depends on the degree to which it is used for comprehension. When L2 fluency is low, explicit grammatical knowledge may be too difficult to access, resulting in underspecification of structure (Clahsen & Felser, 2006). In contrast, when L2 fluency is high, language processing may approach native-like performance, thus reducing reliance on explicit knowledge. Therefore, the metalinguistic buffer effect may occur only in relatively proficient bilinguals, for whom there is a measurable fluency gap between their first and second language. If this conjecture is correct, the language-by-load manipulations deployed in this thesis may provide a unique behavioural means of assessing L2 proficiency and for determining the degree to which L2 processing has transferred from declarative memory to the native linguistic system.

Conclusions

The good-enough approach to language processing has provided a framework for investigating how language inputs may trigger systematic misanalyses that produce off-line errors. The strength of this approach is that it challenges language researchers to reliably demonstrate that on-line processing effects, such as RT latencies or event related potentials, are in fact associated with the comprehension processes that are often presupposed. That is, the claim that an on-line processing effect corresponds with some kind of repair or resolution process may be justified only when comprehension *per se* can also be verified. Otherwise, it may be prudent to assume that the effect corresponds with the detection of a conflict or anomaly, but not necessarily with its resolution. The weakness of the good-enough approach, however, is that it appears to have been too enthusiastic in casting errors of linguistic judgment as errors of comprehension. In turn, the dual-stream architecture it envisions is too narrowly focused on the substrates of the language comprehension system, omitting entirely from the realm of possibility how higher level thought processes might be recruited for off-line judgments. Consequently, the good-enough approach does not sufficiently consider that these higher-level information processes may corrupt retroactively the outputs of the linguistic system. This thesis has illustrated that when the broader cognitive system is taken into account, what is sometimes treated as a comprehension error may in fact reflect an executive control failure, suggesting that even when language comprehension is *good*, it is human judgment that may be merely *good enough*.

References

Altmann, G.T.M., & Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33,* 583-609.

Bader, M., & Meng, M. (2016). Does comprehension (sometimes) go wrong for noncanonical sentences? *Architectures and Mechanisms of Language Processing, Bilbao, Spain.*

Barton, S.B., & Sandford, A.J. (1993) A case study of anomaly detection: shallow semantic processing and cohesion establishment. *Memory & Cognition, 21,* 477-87.

Bates, D., Maechler, M., & Bolker, B. (2013). Lme4: Linear mixed effects models using S4 classes. R package version 0.9999099-2. http://CRAN.R-project.org/package=lme4.

Bever, T.G. (1970). The cognitive basis for linguistic structures. In R. Hayes (Ed.), *Cognition and Language Development* (pp. 279-362). New York: Wiley & Sons, Inc.

Bialystok, E., Craik, F.I.M., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Cell, 16,* 240-50.

Bley-Vroman, R. (1990). The logical problem of foreign language learning. *Linguistic Analysis, 20,* 3-47.

Bock, J.K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology, Learning, Memory, & Cognition, 4,* 575-586.

Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review, 113,* 787-821.

Bortfeld, H., Leon, S.D., Bloom, J.E., Schober, M.F., & Brennan, S.E. (2001). Disfluency rates in conversation: Effects of age, Relationship, topic, role, and gender. *Language and Speech, 44,* 123-47.

Caramazza, A. & Zurif, E.B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language, 3,* 572-582.

Carrithers, C. (1989). Syntactic complexity does not necessarily make sentences harder to understand. *Journal of Psycholinguistic Research, 18,* 75-88.

Carrithers, C., & Bever, T.G., (1984). Eye-fixation patterns during reading confirm theories of language comprehension. *Cognitive Science, 8,* 157-72.

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). Psyscope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, and Computers, 25,* 257-271. http://psy.cns.sissa.it

Chipere, N. (1998). Real language users. Available from: http://cog-prints.ecs.soton.ac.uk/archive/00000712/.

Chomsky, N. (1957). Syntactic Structures. The Hague: Mouton.

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology, 69,* 817-28.

Christianson, K., Hollingworth, A., Halliwell, J.F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology, 42,* 368-407.

Christianson, K., Luke, S.G., & Ferreira, F. (2010). Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 538-544.

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

Clark, H.H, & Clark, E.V. (1968). Semantic distinctions and memory for complex sentences. *Quarterly Journal of Experimental Psychology, 20,* 129-38.

Conway, A.R.A., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, 0., & Engle, R.W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12,* 769-86.

de Almeida, R.G., & Manouilidou, C.M. (2015). The study of verbs in cognitive science. In R.G. de Almeida & C.M. Manouilidou (Eds.), *Cognitive Science Perspectives on Verb Representation and Processing.* New York, NY: Springer.

Dabrawska E., & Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. *Language Sciences, 28,* 604-15.

Dekydspotter, L., Schwartz, B.D., & Sprouse, R.A. (2006). The comparative fallacy in L2 processing research. *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference.* Sommerville, MA.

Erickson, T.A., & Mattson, M.E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal and Verbal Behavior, 20,* 540-22.

Evans, J.S.B.T., & Stanovitch, K.E. (2013). Dual-process theories of higher-cognition: Advancing the debate. *Perspectives on Psychological Science, 8,* 223-241.

Fallows, J. (2015, February 19). 'Mistakes were made'. *The Atlantic.* Retrieved from: https://www.theatlantic.com/politics/archive/2015/02/mistakes-were-made/385663/

Fedorenko, E. (2014). The role of domain-general cognitive control in language comprehension. *Frontiers in Psychology, 5,* 1-17.

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: evidence against domain-specific resources. *Journal of Memory and Language, 54,* 541-553.

Fedorenko, E., & Thompson-Schill, S.L. (2014). Reworking the language network. *Trends in Cognitive Science, 18,* 120-26.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164–203.

Ferreira, F., & Bailey, K. G. D. (2004). Disfluencies and human language comprehension. *Trends in Cognitive Sciences*, *8*, 231–7.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*, 11–15

Ferreira, F., & Lowder, M.W. (2016). Prediction, information structure, and good-enough language processing, *Psychology of Learning & Motivation, 65,* 217-47.

Ferreira, F., & Patson, N. D. (2007). The "good enough" approach to language comprehension. *Language and Linguistics Compass*, *1*, 71–83. doi:10.1111/j.1749-818X.2007.00007.x

Fodor, J.A. (1983). The Modularity of the Mind. MIT Press: Cambridge, MA.

Fodor, J.A., Bever, T.G., & Garrett, M.F. (1974). *The Psychology of Language.* New York, NY: McGraw-Hill.

Fodor, J.A., & Garret, M. (1967). Some syntactic determinants of sentential complexity. *Perception & Psychophysics, 2,* 289-96.

Fodor, J.D., & Innoe, A. (2000). Garden path re-analysis: Attach (anyway) and revision as a last resort. In M. De Vincenzi & V. Lombardo (Eds.), *Cross Linguistic Perspectives on Language Processing,* 21-61.

Fodor, J. A. & Lepore, E. (2002). *The Compositionality Papers.* Oxford, UK: Oxford University Press.

Frazier, L. (1987). Sentence processing: A tutorial review. In M Coltheart (Ed.), *Attention and Performance XIII: The Psychology of Reading.* 561-586.

Frazier, L. & Clifton, C. (2015). Without his shirt off he saved the child from almost drowning: interpreting an uncertain input. *Language, Cognition, & Neuroscience, 30,* 635-47.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210. doi:10.1016/0010-0285(82)90008-1

Frazier, L., Taft, L., Roeper, T., & Clifton, C., & Elrich, K. (1984). Parallel structure: a source of facilitation in sentence comprehension. *Memory & Cognition, 12,* 421-430.

Gundel, J.K., & Fretheim, T. (2004). Topic and focus. In L. Horn & G. Ward (Eds.), *The Handbook of Pragmatics.* Molden, MA: Blackwell.

Hale, J.T. (2011). What a rational parser would do. *Cognitive Science, 35,* 399-443.

Hsu, N.S., & Novic, J.M. (2016). Dynamic engagement of cognitive control modulates recovery from misinterpretation during real-time language processing. *Psychological Science, 27,* 572-82.

Johnson-Laird, P.N. (1968). The interpretation of the passive voice. *Quarterly Journal of Experimental Psychology, 20,* 69-73.

Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review, 99,* 122-49.

Karimi, H., & Ferreira, F. (2015). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology, 69,* 1013-40.

Koornneef, A., & Reuland, E. (2016). On the shallow processing (dis)advantage: grammar and economy. *Frontiers in Psychology, 7.*

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106,* 1126-77.

Li, K.Z.H., Lindenberger, U., Freund, A.M., & Baltes, P.B. (2001). Walking while memorizing: age-related differences in compensatory behaviour. *Psychological Science, 12,* 230-37.

Lim, J.H., & Christianson, K. (2013a). Integrating meaning and structure in L1-L2 and L2-L1 translations. *Second Language Research, 29,* 233-56.

Lim, J.H., & Christianson, K. (2013b). Second language sentence processing in reading for comprehension and translation. *Bilingualism: Language & Cognition, 16,* 518-37.

McMahon, L.E. (1963). Grammatical analysis as part of understanding a sentence. Unpublished Doctoral Thesis.

McWhinney, B., Bates, E., & Kleigl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior, 23,* 127-50.

Mehler, J. (1963). Some effects on grammatical transformations on the recall of passive sentences. *Journal of Verbal Learning and Verbal Behavior, 2,* 346-51.

Miller, G.A. (1962). Some psychological studies of grammar. *American Psychologist, 17,* 748-62.

Miller, G.A., & McKean, K.O. (1964). A chronometric study of some relations between sentences. *Quarterly Journal of Experimental Psychology, 16,* 297-308.

Morton, J. (1966). Discussion of 'On hearing sentences'. In J. Lyons & R.J. Wallers (Eds.). *Psycholinguistics Papers.* Edinburgh: Edinburgh University Press.

Paradis, M. (2004). *A Neurolinguistics Theory of Bilingualism.* Philadelphia: John Benjamins.

Paradis, M. (2009). *Declarative and Procedural Determinants of Second Languages.* Philadelphia: John Benjamins.

Perfetti, C.A., & Goldman, S.R. (1975). Discourse functions of thematization and topicalization. *Journal of Psycholinguistic Research, 4,* 257-71.

Pinker, S. (2014). *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century.* New York, NY: Penguin.

R Development Core Team. (2012). R: A language and environment for statistical computing. *R-Foundation for Statistical Computing*, Vienna, Austria. http://www.R-project.org/

Roland, D., Dick, F., & Elman, J. (2007). Frequency of basic English grammatical structures: a corpus analysis. *Journal of Memory and Language, 57,* 348-79.

Sanford, A., & Sturt, P. (2002). Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Science*, *6*, 382- 386.

Sasaki, Y. (1998). Individual variation in a Japanese sentence comprehension task: form, functions, and strategies. *Applied Linguistics, 18,* 509-37.

Segalowitz, N. (2010). *Cognitive Basis of Second Language Fluency*. New York, NY: Routledge.

Segalowitz, N., & Frenkiel-Fishman, S. (2005). Attention control and ability level in a complex cognitive skill: attention shifting and second-language proficiency. *Memory & Cognition, 33,* 644-53.

Shffrin, R.M., & Schneider, W. (1977). Controlled and automatic human processing: perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127-190.

Slattery, T.J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden paths arise from competing syntactic representations. *Journal of Memory and Language, 69,* 104-20.

Slobin, D.I. (1966). Grammatical transformations and sentence comprehension in childhood and adulthood. *Journal of Verbal Learning and Verbal Behavior, 5,* 219-27.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: evidence from self-paced reading. *Memory & Cognition, 36,* 201-216.

Tanenhaus, M.K., Carlson, G., & Trueswell, J.C. (1989). The role of thematic structures in interpretation and parsing. *Language and Cognitive Processes, 4,* 211-234.

Townsend, D.J., & Bever, T.G. (2001). *Sentence Comprehension: The Integration of Habits and Rules.* Cambridge, MA: MIT Press.

Tolmin, R.S. (1983). On the interaction of syntactic subject, thematic information, and agent in English. *Journal of Pragmatics, 7,* 411-32.

Trueswell, J. C., & Tanenhaus, M.K. (2005). *Approaches to Studying World-Situated Language Use.* Cambridge, MA: MIT Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185,* 1124-31.

Ullman, M.T. (2001). The neural basis of lexicon and grammar in first and second language: the declarative/procedural model. *Bilingualism: Language & Cognition, 4,* 105-122.

Ullman, M.T. (2015). The declarative/procedural model: a neurobiological motivated theory of first and second language. In B. VanPatten & J. Williams (Eds.), *The Neurobiology of Language.* San Diego, CA: Elsevier.

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *Quarterly Journal of Experimental Psychology,* 31, 591–597.

Ye, Z., & Zhou, X. (2009). Executive control in language processing. *Neuroscience and Behavioral Reviews, 33,* 1168-77.

Appendix A

Materials

Active Plausible Sentences

Item	English (Ferreira, 2003)	French
1	The dog bit the man	Le chien a mordu l'homme
2	The cook ruined the food	Le chef a gâché le plat
3	The bird ate the worm	L'oiseau a mangé le ver de terre
4	The cat chased the mouse	Le chat a poursuivi la souris
5	The soldier protected the villager	Le soldat a protégé le villageois
6	The lawyer sued the doctor	L'avocat a poursuivi le médecin
7	The teacher quizzed the student	Le professeur a interrogé l'étudiant
8	The cop pursued the thief	Le policier a poursuivi le voleur
9	The waitress served the man	La serveuse a servi l'homme
10	The owner fed the cat	Le maître a nourri le chat
11	The detective investigated the suspect	Le détective a enquêté sur le suspect
12	The doctor treated the patient	Le médecin a soigné le patient
13	The politician deceived the voter	Le politicien a dupé l'électeur
14	The hiker killed the mosquito	Le randonneur a tué le moustique
15	The horse threw the rider	Le cheval a éjecté le coureur
16	The golfer hit the ball	Le golfeur a frappé la balle
17	The hunter shot the deer	Le chasseur a piégé le lapin
18	The frog ate the fly	La grenouille a mangé la mouche
19	The ghost scared the boy	Le fantôme a effrayé le garçon
20	The horse kicked the jockey	Le cheval a heurté le jockey
21	The angler caught the fish	Le pêcheur a attrapé le poisson
22	The matador dodged the bull	Le matador a esquivé le taureau
23	The officer arrested the citizen	Le policier a arrêté le citoyen
24	The prince slayed the dragon	Le prince a abattu le dragon

Active Implausible Sentences

Item	English Sentences (Ferreira, 2003)	French Sentences
1	The man bit the dog	L'homme a mordu le chien
2	The food ruined the cook	Le plat a gâché le chef
3	The worm ate the bird	Le ver de terre a mangé l'oiseau
4	The mouse chased the cat	La souris a poursuivi le chat
5	The villager protected the soldier	Le villageois a protégé le soldat
6	The doctor sued the lawyer	Le médecin a poursuivi l'avocat
7	The student quizzed the teacher	L'étudiant a interrogé le professeur
8	The thief pursued the cop	Le voleur a poursuivi le policier
9	The man served the waitress	L'homme a servi la serveuse
10	The cat fed the owner	Le chat a nourri le maître
11	The suspect investigated the detective	Le suspect a enquêté sur le détective
12	The patient treated the doctor	Le patient a soigné le médecin
13	The voter deceived the politician	L'électeur a dupé le politicien
14	The mosquito killed the hiker	Le moustique a tué le randonneur
15	The rider threw the horse	Le coureur a éjecté le cheval
16	The ball hit the golfer	La balle a frappé le golfeur
17	The deer shot the hunter	Le lapin a piégé le chasseur
18	The fly ate the frog	La mouche a mangé la grenouille
19	The boy scared the ghost	Le garçon a effrayé le fantôme
20	The jockey kicked the horse	Le jockey a heurté le cheval
21	The fish caught the angler	Le poisson a attrapé le pêcheur
22	The bull dodged the matador	Le taureau a esquivé le matador
23	The citizen arrested the officer	Le citoyen a arrêté le policier
24	The dragon slayed the prince	Le dragon a abattu le prince

Passive Plausible Sentences

Item	English Sentences (Ferreira, 2003)	French Sentences
1	The man was bitten by the dog	L'homme a été mordu par le chien
2	The food was ruined by the cook	Le plat a été gâché par le chef
3	The worm was eaten by the bird	Le ver de terre a été mangé par l'oiseau
4	The mouse was chased by the cat	La souris a été poursuivie par le chat
5	The villager was protected by the soldier	Le villageois a été protégé par le soldat
6	The doctor was sued by the lawyer	Le médecin a été poursuivi par l'avocat
7	The student was quizzed by the teacher	L'étudiant a été interrogé par le professeur
8	The thief was pursued by the cop	Le voleur a été poursuivi par le policier
9	The man was served by the waitress	L'homme a été servi par la serveuse
10	The cat was fed by the owner	Le chat a été nourri par le maître
11	The suspect was investigated by the detective	Le suspect a été enquêté par le détective
12	The patient was treated by the doctor	Le patient a été soigné par le médecin
13	The voter was deceived by the politician	L'électeur a été dupé par le politicien
14	The mosquito was killed by the hiker	Le moustique a été tué par le randonneur
15	The rider was thrown by the horse	Le coureur a été éjecté par le cheval
16	The ball was hit by the golfer	La balle a été frappée par le golfeur
17	The deer was shot by the hunter	Le lapin a été piégé par le chasseur
18	The fly was eaten by the frog	La mouche a été mangée par la grenouille
19	The boy was scared by the ghost	Le garçon a été effrayé par le fantôme
20	The jockey was kicked by the horse	Le jockey a été heurté par le cheval
21	The fish was caught by the angler	Le poisson a été attrapé par le pêcheur
22	The bull was dodged by the matador	Le taureau a été esquivé par le matador
23	The citizen was arrested by the officer	Le citoyen a été arrêté par le policier
24	The dragon was slayed by the prince	Le dragon a été abattu par le prince

Passive Implausible Sentences

Item	English Sentences (Ferreira, 2003)	French Sentences
1	The dog was bitten by the man	Le chien a été mordu par l’homme
2	The cook was ruined by the food	Le chef a été gâché par le plat
3	The bird was eaten by the worm	L’oiseau a été mangé par le ver de terre
4	The cat was chased by the mouse	Le chat a été poursuivi par la souris
5	The soldier was protected by the villager	Le soldat a été protégé par le villageois
6	The lawyer was sued by the doctor	L’avocat a été poursuivi par le médecin
7	The teacher was quizzed by the student	Le professeur a été interrogé par l’étudiant
8	The cop was pursued by the thief	Le policier a été poursuivi par le voleur
9	The waitress was served by the man	La serveuse a été servie par l’homme
10	The owner was fed by the cat	Le maître a été nourri par le chat
11	The detective was investigated by the suspect	Le détective a été enquêté par le suspect
12	The doctor was treated by the patient	Le médecin a été soigné par le patient
13	The politician was deceived by the voter	Le politicien a été dupé par l’électeur
14	The hiker was killed by the mosquito	Le randonneur a été tué par le moustique
15	The horse was thrown by the rider	Le cheval a été éjecté par le coureur
16	The golfer was hit by the ball	Le golfeur a été frappé par la balle
17	The hunter was shot by the deer	Le chasseur a été piégé par le lapin
18	The frog was eaten by the fly	La grenouille a été mangée par la mouche
19	The ghost was scared by the boy	Le fantôme a été effrayé par le garçon
20	The horse was kicked by the jockey	Le cheval a été heurté par le jockey
21	The angler was caught by the fish	Le pêcheur a été attrapé par le poisson
22	The matador was dodged by the bull	Le matador a été esquivé par le taureau
23	The officer was arrested by the citizen	Le policier a été arrêté par le citoyen
24	The prince was slayed by the dragon	Le prince a été abattu par le dragon

Appendix B

Below are the results of a mixed-effects logistic regression model performed with the data from Manuscript 2, including Blocks 7 and 8, which were removed to minimize practice effects. The results reported below reflect the same analyses reported in the Accuracy section of Manuscript 2 (p.57-59), and produced the same results, with minor variation in the magnitude of effects. Table 12 below shows the statistically significant estimates obtained in the analysis. All other effects were not significant at the conventional threshold (max z=|-1.86|, min p=0.062). As in Manuscript 2 (as well as previous studies), there was a main effect of voice and plausibility. The odds of correct agentive assignment in the active voice were 62 times that of the passive voice. Similarly, the odds of correct agentive assignment in the plausible conditions were 2.3 times that of the implausible conditions. Most importantly, the language-by-load crossover interaction, which represents the main result of Manuscript 2, was also found for the model below. While there was no main effect of language, z=-1.86, p=.062, or load, z=1.11, p=0.266, there was a significant crossover interaction between these factors conforming to the pattern reported in Manuscript 2: in the single task condition, bilinguals reported more correct responses in their L2 , but in the dual task condition, they reported more correct responses in their L1. The main divergence between the present model and that of Manuscript 2, was that here, we also obtained a voice-by-language interaction, such that the effect of the passive voice was more pronounced in the L1 condition, but the numerical differences were in fact quite marginal, with an active-passive mean difference of 26.5% in the L1 condition and 24.3% in the L2 condition. These small differences are underscored by the fact that there was no main effect of language, due to the crossover effects associated with the load factor.

Table 12

Mixed-effects logistic regression model of thematic judgment accuracy including all 8 blocks from Manuscript 2.

Fixed Effects	E	SE	z	*p*	*OR* (e^{E})	95% CI
Voice	-4.13	0.39	-10.49	< 0.001	0.02	[0.01, 0.03]
Plausibility	0.96	0.42	2.27	0.023	2.60	[1.14, 5.95]
Voice * Language	0.94	0.39	2.40	0.016	2.57	[1.19, 5.56]
Load * Language	-0.66	0.23	-2.82	0.005	0.52	[0.33, 0.82]