Sentence Processing in Relative Clauses in Standard Arabic

By

NICOLE DODD

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Linguistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____

Emily Morgan, Chair

_____

Fernanda Ferreira

_____

Masoud Jasbi

Committee in Charge

2024

Abstract

What makes some sentences more difficult to process, and why? Memory- and expectation-based theories both attempt to explain sentence processing difficulties, and decades of sentence processing literature have found evidence in support of both theories. This dissertation further investigates these theories of sentence processing by exploring processing of subject- and object-extracted relative clauses (SRCs and ORCs) in Modern Standard Arabic, and how expectations affect the resulting interpretation of low-frequency structures. We investigate this question through various experimental paradigms. We first tested memory- versus expectation-based theories using a self-paced reading task. Results showed longer reading times for ORCs, supporting expectation-based theories, with difficulty localized at the relative clause noun phrase. We also found that misinterpretations were more frequent for ORCs, suggesting possible misreading as SRCs or good-enough and noisy-channel processing. To investigate this phenomenon, we conducted a recall task where participants re-wrote sentences word-for-word. Errors showed both ORCs being re-written as SRCs and vice versa, supporting good-enough and noisy-channel processing theories. We then explicitly tested the possibility of misreading versus good-enough or noisy-channel processing through eye-tracking. Findings indicated that readers were not misreading ORCs, but instead accepting noisy SRC interpretations. Finally, we explored the impact of grammatical cues versus statistical expectations in a second eye-tracking experiment. Results showed that increasing grammatical cues in favor of a veridical ORC interpretation did not significantly affect noisy interpretations or processing behaviors, indicating that grammatical cues were insufficient to override statistical expectations in a good-enough or noisy-channel framework.

Acknowledgements

Completing a PhD is seen as a highly individualistic accomplishment, and while to a certain extent that is true, there is a mountain of people without whom I would never have reached the finish line.

Thank you first and foremost to my advisor, Emily Morgan. Thank you for supporting my research in a language you knew little about, and for helping me find my academic voice. I am grateful for the mentorship and guidance you provided along the way, and for modeling a healthy work-life balance in an industry that so often overlooks such balances. I am honored to be your first PhD graduate.

Thank you also to my committee members, Fernanda Ferreira and Masoud Jasbi. Fernanda, thank you for providing valuable insight into my research for my whole program and for playing a pivotal role in securing the funding needed to make my research a success. Masoud, thank you for providing much-needed perspectives on the Arabic language and for providing valuable feedback on both my preliminary paper and now my dissertation.

I am wholly indebted to the many native speakers without whom I could not have completed this research: Manar Al-Shatarat, Shayma Hassouna, Mariam Ali, Reem Suleiman, Meera Al-Kaabi, Fatima Boush, and many others that I spoke to at HSP, CogSci, and beyond. Your insight into the Arabic language was crucial for this non-native speaker that took on the behemoth task of studying a language she does not natively speak.

I am also incredibly grateful for the team at the United Arab Emirates University – Meera Al-Kaabi, Tommi Leung, and Fatima Boush – who invited me to collect data on their campus for a month, and assisted with many logistics before and after data collection. The data collected at the UAEU was paramount for my research program. Thank you also to the National Science Foundation for funding this field research, which was the deciding factor in my ability to go.

Graduate school can feel very isolating at times. I am thankful for the many colleagues who helped me feel a little less alone along the way: the MA crew in our overloaded closet office, the colleagues who

became roommates and friends, my small but mighty cohort, those who always stopped to chat in the hall, and the many others I met in Davis' grad community. Thank you for providing support and understanding in a way that few outside our experience can.

So many friends and family provided support from behind the scenes these past five years. Mom and Dad, thank you for always supporting my next endeavor no matter how much it may surprise you. Kayla and Kira, thank you for being my fellow doctor friends (though in a very different way!). Rachel, thank you for always being a listening ear from thousands of miles away. Nicole (which one?), thank you for always hyping up Dr. Dodd. Kenzie, thanks for understanding my research better than most. And thank you to the many others I can't exhaustively list here – you know who you are.

Finally, thank you to my husband John, who came into my life when I least expected it and was the least prepared for it. Thank you for believing in me, encouraging me, and loving me. You keep me grounded on the days that I have academia tunnel vision and remind me that there is more to my identity than this. I love you.

For Grandpa John

Declarations

This research was supported by the National Science Foundation Doctoral Dissertation Research Improvement Grant to ND, EM, and FF [grant number 2235106], the UC Davis Society of Hellman Fellows Program to EM, and the UC Davis College of Letters and Science to EM. The Penn Arabic Treebank Part 3 v 3.2 was provided by the Linguistic Data Consortium through a Data Scholarship to ND.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

In the cognitive science of language, one main question is how people understand whole sentences. For a fluent speaker of a language, reading typically feels effortless. But in fact, some sentences are more difficult to read and understand than others. For example, sentences with more common structures like (a) "The reporter who attacked the senator admitted the error" are easier to process than those with less common structures, like (b) "The reporter who the senator attacked admitted the error." People expect to see (a) since it's more common, but if they instead see (b), they have a harder time processing it because it contradicted their expectations. By monitoring people's eye movements while they read sentences like (b) versus (a), we see that people spend more time reading and are more likely to re-read sentence (b), two signs of processing difficulty.

Reading difficult sentences can also lead to misinterpretations. For instance, someone may read sentence (b) and understand that the reporter attacked the senator, when in reality, the senator attacked the reporter. We can imagine two possible reasons for this. They may have simply misread it – there's only a subtle change in the word order between (a) and (b), but we have to notice it to get the correct interpretation. On the other hand, there's evidence that people sometimes initially read a sentence correctly but then subconsciously change their beliefs about what they read, perhaps because (a)'s structure is far more common than (b).

My research investigated how people read and understand difficult sentences, using Arabic as a case study. I measured people's reading times and monitored reader's eye movements as they read Arabic versions of sentences like (a) and (b), and measured the accuracy of their interpretations by having them answer comprehension questions after each sentence. I also tested whether misinterpretations were due to misreading a sentence or accepting a "good-enough" interpretation of the sentence. Finally, I investigated the tradeoff of expectations and input during processing, and how readers make use of expectations in light of conflicting input. The findings from this research contribute to our understanding of language processing phenomena by providing further evidence for memory- and expectation-based sentence processing theories, and elucidating the relationships between the various factors that contribute to good-enough or noisy-channel processing.

1.1. Memory- and expectation-based theories of sentence processing

What makes some sentences more difficult to read and comprehend than others? Two main types of theories aim to explain causes of sentence processing difficulty: memory-based theories (Gibson, 1998; Gibson, 2000; Grodner & Gibson, 2005) and expectation- or constraint-based theories (Hale, 2001; MacDonald et al., 1994; Levy, 2008b). Memory-based theories claim that processing difficulties arise while processing structures that require a large amount of our limited cognitive computational resources, while expectation-based theories claim that processing sentences that do not align with our syntactic or semantic expectations can incur additional processing costs. These theories are often tested cross-linguistically using subject- and object-extracted relative clauses (Lau & Tanaka, 2021). In subject-extracted relative clauses (SRCs), the noun phrase (NP) subject of the matrix clause is also the subject of the relative clause; in object-extracted relative clauses (ORCs), the NP subject of the matrix clause is the object of the relative clause (Figure 1.1).

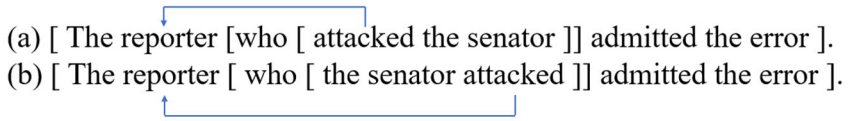

Figure 1.1: Example (a) SRC and (b) ORC in English. Dependencies between the relative clause verb and matrix clause subject are shown in blue.

Memory-based theories predict more processing difficulty when reading structures that utilize more working memory during incremental processing. One example of how this difficulty presents is with sentences with long distance dependencies between constituents. Humans have limited computational resources, so readers incur more processing costs the longer they maintain structures with incomplete dependencies in memory. For example, in English relative clauses, the dependency from the relative clause verb to the matrix noun is longer for ORCs than for SRCs (Figure 1.1). This results in more demands on working memory while reading ORCs, and thus more processing difficulty. Further, an additional cost is paid upon integrating the long dependency with the existing structure of the sentence (e.g., integrating "the reporter" upon resolving the dependency at "attacked"). This memory-based processing difficulty is formalized in the Dependency Locality Theory (Gibson, 2000), which states that the cost of processing and integrating two elements is directly proportional to the length of the dependency between the elements.

Expectation-based theories posit that items that are less expected or lower frequency in context are more difficult to process. During incremental processing, all possible syntactic parses are evaluated in parallel, and potential parses are eliminated as more contextual information becomes available. Potential parses are preferentially ranked given contextual cues, and processing difficulty arises when a low-ranked parse is the resulting correct structure; thus, expectation-based processing difficulty arises when the reader encounters an element that violates their expectations for the upcoming syntactic parse. The reader then pays a processing cost proportional to the difficulty of updating their expectations. Early iterations of expectation-based theories, such as constraint-based theories, stated that possible syntactic parses were

constrained by the interaction of contextual factors such as lexical and syntactic frequency, thematic roles, and argument structures (MacDonald et al., 1994; Spivey-Knowlton & Tanenhaus, 1998). Frequency-based tuning models make a similar claim, stating that language processing is highly dependent on input frequency across a variety of linguistic factors (i.e., phonological frequency, lexical frequency, etc.) (e.g., Ellis, 2002). Building upon these constraints, more recent expectation-based theories have modeled these contextual expectations using surprisal theory (i.e., negative log-probability of a word given previous context) (Hale, 2001; Levy, 2008b).

In English, SRCs are more common and thus more expected than ORCs (Roland et al., 2007). When reading an ORC, readers will incur processing difficulty after reading the relative pronoun "who", where the reader expects to encounter a verb (e.g., "attacked"), signaling an SRC, but instead encounters an NP (e.g., "the senator"), signaling an ORC (Figure 1.1). Many expectation-based theories operationalize this cost using surprisal theory, calculated as the negative log-probability of a word given previous context (Hale, 2001; Levy, 2008b). Words with a larger surprisal value are more surprising in context and are therefore predicted to be read more slowly than words with smaller surprisal values.

Some studies testing these two classes of theories have found evidence directly in support of and in contradiction to one theory. For example, Konieczny & Döring (2003) tested memory- and expectation-based theories of sentence processing in German verb-final clauses. The authors compared processing times for verb-final clauses with one manipulation: in one sentence, the second clausal noun functioned as the indirect object of the clause (dative case), and in the other sentence, it functioned as a descriptor for the main clause noun (genitive case). As each clause was head-final, all processing costs from integrating the syntactic parse and resolving incomplete dependencies were paid upon encountering the clause-final verb. The authors predicted that clauses that introduced a higher number of dependents for the verb before encountering the clause-final verb would have faster processing times. In accordance with expectation-based theories, they argued that each additional dependent to the verb helped to narrow the number of possible syntactic parses being processed in parallel, and this would result in faster processing times at the

final point of integration. The findings showed that the dative construction resulted in faster processing times at the clause-final verb, and thus provided support for expectation-based theories and contradicted predictions from memory-based theories in German language processing.

On the other hand, other research has found that both memory- and expectation-based constraints can contribute to processing difficulties. For example, in English, memory- and expectation-based theories both predict increased processing difficulties in ORCs: memory-based theories predict faster processing with SRCs due to the shorter dependency between the relative clause verb and its dependent, the relative pronoun, and expectation-based theories predict faster processing with SRCs as it is the higher frequency structure. Crucially, these two theories predict different loci of processing difficulty. Expectation-based theories predict processing difficulty upon encountering the relative clause noun (e.g., "the senator), as encountering a noun in the relative clause prior to the verb violates expectations for the structural parse. Meanwhile, memory-based theories predict processing difficulty upon encountering the clausal verb (e.g., "attacked"), as this is where the long-distance dependency between the relative pronoun and the clausal verb is resolved and the integration cost is paid. Staub (2010) conducted an eye tracking study with English relative clauses and found that *both* memory-based constraints and expectations contributed to processing difficulties. Notably, these difficulties manifested in distinct behaviors: difficulty due to memory constraints presented as longer go-past reading times while difficulty due to violated expectations presented as an increased number of regressive saccades. Staub thus argued that both theories explain observed processing difficulty, and each may be tied to a specific processing behavior.

Given these findings, recent work has aimed to develop computational models that represent the joint contributions of memory- and expectation-based costs to language processing. One example of this is the Psycholinguistically Motivated Tree-Adjoining Grammar (PLTAG; Demberg & Keller, 2008, 2009; Demberg, Keller, & Koller, 2013). PLTAG is a computational parsing model built upon previous versions of tree-adjoining grammar that incorporates the structural connectedness and probabilistic prediction components of incremental human processing. The model represents processing costs as predicted by

memory- and expectation-based theories through the use of a probabilistic parser and a prediction and verification mechanism that accounts for memory decay at the point of integration. The authors compared the probabilistic outcomes from the parser to previous findings in processing literature and found that the model generated predictions that accurately modeled patterns of processing difficulty from previous studies.

While this model accurately represented the independent contributions of each processing difficulty, other models have attempted to represent both the independent and interactive contributions of these effects. One such example of this type of model is the Noisy-Context Surprisal model (Futrell et al., 2020; Futrell & Levy, 2017; Hahn et al., 2022). This model measures the processing cost of a word as a function of its surprisal given a noisy representation of the previous context. The surprisal of a word in context directly represents predictions from expectation-based processing theories (low surprisal = high frequency), and the noisy representation models the degradation of human memory during incremental sentence processing. This model additionally introduces the information locality theory, a derivation of DLT, as an explanation for costs incurred from memory-based constraints. While DLT states that constituents with shorter dependencies are easier to process, information locality theory asserts that collocated words with higher levels of mutual information are easier to process. This use of information locality theory updates predictions from memory- and DLT-based constraints theories to include probabilistic expectations. A more recent version of this model, the Lossy-Context Surprisal model (Futrell et al., 2020), replaces the original noisy-context parameter with a lossy-memory representation of context, which more accurately represents the degradation of memory and its effect on incremental processing.

Overall, researchers now largely agree that both memory constraints and expectations contribute to processing difficulty and are not inherently contradictory theories; rather, each constraint can contribute distinct processing difficulty while reading, and in some cases, dictate that these difficulties will be encountered at different parts of the sentence. What remains to be answered in this space is exactly how

these two constraints interact: whether one is stronger than the other, whether they show up as distinctive reading behaviors (e.g., Staub, 2010), and how language-specific differences affect when and how these difficulties present.

1.2. Misinterpretations while reading

Violated expectations during reading can both cause increased processing difficulty and result in the misinterpretation of a sentence. This is the case in models of good-enough (Ferreira et al., 2002; Ferreira & Lowder, 2016; Ferreira & Patson, 2007; Huang & Ferreira, 2021) and noisy-channel processing (Gibson et al., 2013; Keshev & Meltzer-Asscher, 2021; Levy, 2008a; Levy, 2011; Levy et al., 2009, but cf. Cutter et al., 2022; Poppels & Levy, 2016).

Models of good-enough processing state that readers often construct superficial representations of input during processing that may nonetheless be "good enough" to support communication goals. Errors in language comprehension can then occur when readers fail to appropriately access lexical or grammatical constructions. Readers may reanalyze the input when their initial representation is incorrect, but the lingering incorrect interpretation can interfere with arriving at the correct meaning of a sentence. In the case of syntactic processing, encountering a structure that is unexpected in context may cause a reader to reanalyze, yet still accept the interfering "good-enough" interpretation of the input.

One example of this is evident in the interpretation of garden path sentences (e.g., "While Bill hunted the deer ran into the woods," where readers are likely to initially incorrectly assign the thematic role of direct object to "the deer" – "Bill hunted the deer" – until they encounter the second verb "ran," which violates this initial syntactic parse). For example, Christianson et al. (2001) presented participants with garden path sentences of varying lengths and asked participants comprehension questions related to the initial parse (e.g., "Did Bill hunt the deer?"). They found that participants were significantly likely to incorrectly respond "Yes, Bill hunted the deer" after reading a garden path sentence, suggesting that the initial incorrect interpretation lingered and interfered with the correct interpretation.

Noisy-channel processing makes a similar claim, operationalized through statistical reasoning. Language input takes place in noisy circumstances – such as human error and competing environmental conditions – and this noise affects language processing strategies. Noisy-channel processing theories thus suggest that language users weigh the probability of a given sentence structure against the probability of noisy input during sentence processing. In cases where different syntactic structures are possible but one is higher probability than the other, a reader may assume noise in the input and make a number of "edits" to a sentence to arrive at the higher-probability interpretation (e.g., Keshev & Meltzer-Asscher, 2021; Levy, 2011; Poppels & Levy, 2016).

For example, Levy et al. (2009) conducted an eye tracking experiment to investigate whether readers maintain word-level uncertainty during syntactic processing. They used sentences with near-lexical neighbor words, where the reader could make a simple "edit" and interpret the sentence to have a different syntactic structure (e.g., "The coach smiled at the player tossed the frisbee" where "at" could be confused with "as" or "and" and result in a different thematic role for "the player", versus "The coach smiled toward the player tossed the frisbee," where "toward" does not have any such near neighbors). Their results showed that readers were significantly more likely to incorrectly interpret the sentences with near neighbor words than those without. Further, they found that readers were more likely to make a regressive saccade from the critical verb ("tossed") and were more likely to regress back to the preposition upon fixating the verb in the near-neighbor condition. They take these results to be evidence that (a) readers maintain some level of word-level uncertainty during processing, and (b) readers engage in rational probabilistic inference during processing, which in turn affects reading strategies. Gibson et al. (2013) further tested this paradigm by focusing on expectations based on semantic cues. They generated pairs of sentences that were grammatically correct but semantically implausible and required various numbers and types of edits to arrive at the correct interpretation. For example, "the girl kicked the ball" is semantically plausible while "the girl was kicked by the ball" is grammatical but semantically implausible, and would require two edits – specifically, two deletions ("was" and "by") – to become

plausible. The authors found that the probability of an implausible sentence being noisily interpreted as a plausible one was dependent on the number and type of edits it would take to arrive at the plausible interpretation: sentences that required fewer edits to arrive at the plausible interpretation had higher rates of noisy interpretations than those that required more edits, and sentences that required deletions had higher rates of noisy interpretations than sentences that required insertions.

In summary, both theories predict that readers experience increased processing difficulty when encountering violated expectations, and may accept the wrong, but more probable, interpretation of the sentence. Since both good-enough and noisy-channel processing theories make similar predictions, we do not attempt to differentiate between them here and consider them jointly.

1.3. Case study: subject- and object-extracted relative clauses

Subject- and object-extracted relative clauses have long been a popular case study for investigating theories of sentence processing as nearly all the world's languages include this type of structure (Lau & Tanaka, 2021). Early studies that investigated processing differences in SRCs and ORCs cross-linguistically suggested a "universal" subject advantage while processing relative clauses. This effect was based on evidence from primarily Indo-European languages, such as English (Gordon et al., 2001; King & Just, 1991; Traxler et al., 2002; Traxler et al., 2005), German (Friederici et al., 1998; Schriefers et al., 1995), and Dutch (Mak et al., 2002), as well as some other non-Indo-European languages, such as Japanese (Ueno & Garnsey, 2008) and Korean (Kwon et al., 2013; Kwon et al., 2010). Subsequent studies, however, found that this SR advantage was not universal; for example, in Chinese (Chen et al., 2008; Hsiao & Gibson, 2003) and Basque (Carreiras et al., 2010), SRCs were harder to process than ORCs.

The variance among these cross-linguistic findings can often be attributed to differences in typological factors such as word order (e.g., SVO vs. SOV), clause-headedness (head-initial vs. head-final), relative clause positioning in a sentence (pre-nominal vs. post-nominal), and the use of resumptive

pronouns (RPs) (Lau & Tanaka, 2021). While processing patterns in relative clauses are largely language- and feature-specific, languages with different typological features are not evenly represented in previous research. For example, more research has been done on Indo-European languages than languages in other language families, and has studied SVO and SOV word-order languages more than VSO word-order languages. Our research takes steps toward diversifying this body of research by investigating Modern Standard Arabic, a morphosyntactically-complex language that uses both VSO and SVO word orders and is generally under-represented in psycholinguistic literature.

1.4. Modern Standard Arabic

Modern Standard Arabic (MSA) is a Semitic language that is written right-to-left using a continuous cursive script. All letters in the Arabic alphabet are consonants – except three that can additionally be used as long vowels – and short vowels are indicated using diacritic marks above or below a consonant (Holes, 2004). In the absence of short vowels, many Arabic words are ambiguous and must be disambiguated by context (Abu–Rabia, 2001). For example, the consonants /ktb/ (كتب) can mean "he wrote" /kataba/ (كَتَبَ) or "books" /kutubun/ (كُتُبٌ) depending on the short vowels used. However, diacritics are typically only included in religious texts and rarely used in everyday written Arabic, except to occasionally disambiguate words (Hermena et al., 2015).

Arabic demonstrates substantial morphosyntactic complexity. Words are derived using non-concatenative morphology with a trilateral root system (Abu–Rabia, 2002; Boudelaa & Marslen-Wilson, 2001), where inflectional morphemes are inserted between the three root consonants, as well as at the beginning and end of a word. For example, the consonant root /ktb/ (كتب) is the base for words such as "he writes" /jaktubu/ (يكتب), "writing" /kita:ba/ (كتابة), and "writer" /ka:tib/ (كاتب). Arabic also makes extensive use of bound pro- and enclitics, which include conjunctions (e.g., "and" /wa=/ (و) – "a book and a pen" /kita:b wa=qalam/ (كتاب وقلم)), prepositions (e.g., "with" /bi=/ (ب) – "he wrote with the pen" /kutab bi=alqalam/ (كتب بالقلم)), definite markers (e.g., "the" /al=/ (ال) – "the book" /al=kita:b/ (الكتاب)), and possessive pronouns (e.g., "her" /=ha/ (ها) – "her daughter" /ibnatu=ha/ (ابنتها)). Because of this extensive

morphology, Arabic words are more information dense than words in Latinate languages of comparable length (Boudelaa & Marslen-Wilson, 2010; Brysbaert, 2019; Roman & Pavard, 1987).

Arabic is a flexible word order language and uses alternating SVO and VSO word order (Parkinson, 1981; Ryding, 2005). Basic word order in MSA tends to be VSO, while SVO is often used in a stylistic manner or to emphasize or topicalize the subject. On the other hand, many dialects of Arabic have basic SVO word order (Parkinson, 1981).

MSA is mainly used in official governmental or media domains, and native Arabic speakers typically learn MSA alongside their regional dialect used for everyday communication. While MSA usage is largely domain-specific, MSA is not considered to be a "second language" for native speakers of Arabic. Previous research has found that native Arabic speakers have comparable proficiencies in MSA and their colloquial dialect; however, notably, proficiency in MSA was strongest in reading and writing compared to speaking, as native Arabic speakers typically read and write in MSA but speak in their colloquial dialect (Albirini, 2019).

1.4.1. Language processing in MSA

While MSA and its regional dialects are the fifth most spoken first language in the world (Central Intelligence Agency, 2018), psycholinguistic research in Arabic is sparse (Hermena, 2016). Much of the previous work in Arabic language processing has focused on word-level processing phenomena related to Arabic's written script, diacritic markings, and morphology in order to identify differences in processing patterns found in other Indo-European languages. One key finding from these studies was how native Arabic speakers fixate words while reading. Farid & Grainger (1996) compared initial fixation positions in Arabic and French found that asymmetrical initial word fixation in Arabic was modulated by prefixation and suffixation in the stimuli. Specifically, fixation patterns revealed a stem bias in which prefixed words tended to have left initial fixation bias, and suffixed words tended to have right initial fixation bias (NB: Arabic is read right to left). The same effects were not found for French: French

exhibited a left initial fixation bias regardless of the affixation of the stimulus. The authors speculated that this pattern of initial fixation was due to the locality of Arabic's information density as a trilateral root language. Speakers of many Indo-European languages benefit from word-initial fixation as it provides information to predict the upcoming word; on the other hand, the majority of word meaning lies at the root in Arabic, and thus this word-initial fixation is not as beneficial for prediction purposes. Later research also identified a distinct perceptual span asymmetry in Arabic reading. Jordan et al. (2014) investigated reading behaviors in native Arabic speakers with strong proficiency in English and found that, while perceptual spans extend asymmetrically to the right in English (Pollatsek et al., 1981), perceptual spans extend asymmetrically to the left in Arabic. These studies provided further evidence that fixation patterns and perceptual span asymmetries are largely a function of reading direction.

Other studies have investigated how word length and frequency affect processing behaviors in Arabic. Paterson et al. (2015) investigated the effects of word length on eye movement patterns by using sentences with target words that were either 3, 5, or 7 characters in length and matched in contextual and lexical frequency. Their results showed that reading times were much longer in Arabic than words of a similar length in Latinate languages like English (Rayner, 2009). The authors theorized that this may be due to the increased information density in Arabic words (Brysbaert, 2019). Another study by Hermena and colleagues tested how frequency affected reading patterns in Arabic (Hermena et al., 2019). They used target words that were either high or low frequency, but matched for both character length and spatial length. They again found that overall reading times were much longer and overall word-skipping rates were low (less than 4%) in Arabic compared to languages like English and German, supporting findings from Paterson et al. (2015). For high-frequency words specifically, they found that readers made fewer fixations and had shorter reading times, which aligns with frequency effects observed in other Latinate languages (e.g., Inhoff & Rayner, 1986; Rayner, 1998, 2009). However, they also observed that skipping rates did not differ across low and high-frequency words, suggesting that Arabic readers are less likely to use information about lexical frequency to skip words.

Sentence-level processing studies have focused on how diacritization and word ordering affect processing difficulty in Arabic. In one experiment, Hermena et al. (2015) explored the effects of diacritization by presenting readers with sentences with no diacritization, sentences that were fully vowelized, and sentences where diacritics were used only to disambiguate specific words. They found that the diacritization of ambiguous words was helpful for processing, but full sentence diacritization slowed overall reading times. The authors attributed this effect to the fact that full vowelization with diacritics increased visual crowding in the text, thus causing longer reading times. These findings also confirm what is expected given Arabic readers' experience with diacritized text: typically, Arabic text is not diacritized in everyday writing. Other research has investigated how readers process the two primary word orders for Arabic – VSO and SVO – and whether one word order is processed faster than another. One study investigated sentence processing in Saudi Arabian speakers, which has SVO basic word order (Thompson & Werfelli, 2012). They found that, despite the fact that their dialect had primarily SVO word order, Saudi Arabian speakers had faster reading times in sentences with VSO word order than sentences in SVO word order. Another study conducted a similar experiment with Saudi Arabian native and heritage speakers and again found a general processing preference for VSO word order (AlQahtani & Sabourin, 2015). However, this effect was slightly modulated by speaker type: native speakers read VSO word order sentences faster across the board, while heritage speakers processed VSO fastest when reading an indefinite subject, but SVO fastest when reading a definite subject.

What remains to be explored in this literature is general patterns of sentence processing that relate back to cross-linguistic theories of sentence processing. Thus, our research aims to both diversify existing language processing literature and augment existing research in Arabic psycholinguistics by taking advantage of the unique features of MSA to distinguish among competing psycholinguistic theories.

1.4.2. Relative clauses in MSA

Restrictive relative clauses in MSA can take a number of forms due to its flexible word order and varying options for marking an object within a relative clause. A sample SRC and ORC is shown in Figure 1.2.

Arabic uses both SVO and VSO; however, VSO is seen as the default word order for MSA and has also been found to be processed faster than SVO (AlQahtani & Sabourin, 2015; Thompson & Werfelli, 2012). This flexible word order means that sentences with embedded relative clauses could appear with four different ordering combinations: the matrix clause in VSO or SVO, and the relative clause in VSO or SVO. We discussed these possible combinations with a handful of native speakers who stated that, while all combinations were grammatical, a SVO matrix clause with a VSO relative clause was the most preferred.

spillover 1	matrix verb	RC NP	RC verb	RC pronoun	matrix NP	
بالخطأ	اعترف	السناتور	هاجم	الذي	الصحفي	(a)
b=il=xaðˤʕ-i	<ʔ>ʕ<ta>rafa	a:=si:na:tu:r	h<a:>ʒam	a:la-ði	a:=sˤaħafi-u	
to=DET=error-ACC	admit<3SG.M.PST>	DET=senator	attack<3SG.M.PST>	who-3SG.M	DET=reporter-NOM	

SRC: "The reporter who attacked the senator admitted the error."

بالخطأ	اعترف	السناتور	هاجمه	الذي	الصحفي	(b)
b=il=xaðˤʕ-i	<ʔ>ʕ<ta>rafa	a:=si:na:tu:r	h<a:>ʒam=ahu	a:la-ði	a:=sˤaħafi-u	
to=DET=error-ACC	admit<3SG.M.PST>	DET=senator	attack<3SG.M.PST>=3SG.M.ACC	who-3SG.M	DET=reporter-NOM	

ORC: "The reporter who the senator attacked admitted the error."

Figure 1.2: Sample (a) SRC and (b) ORC stimuli using VSO relative clause word order and the resumption strategy. Arabic sentences and English glosses are read right to left. The red circles indicate the disambiguating region: the relative clause verb. The only difference between an SRC and ORC in these forms is the presence of a resumptive object pronoun clitic on the relative clause verb in the ORC condition.

MSA also allows both the gap strategy and the resumption strategy when marking direct objects in ORCs, but restricts the use of gapping in certain cases (Aoun et al., 2010). For example, sentences with indefinite relativized noun phrases must use resumption and include a grammaticalized resumptive object pronoun clitic within the relative clause. On the other hand, sentences with definite relativized noun phrases can either mark the object position with a gap or a resumptive object pronoun clitic. Among these

two strategies, resumption is seen as the default strategy in MSA and is required in many dialects of Arabic[1, 2] (Alresaini, 2012, 2016; Aoun et al., 2010; Leung et al., 2020).

Previous research in sentence processing has spent little time investigating languages that use VSO word order and languages that use grammaticalized resumptive pronoun clitics. Our research thus focuses on relative clauses that incorporate these two components to better understand processing behaviors with these linguistic features. With this structure, the reader will read the relative clause verb first for both SRCs and ORCs, and an ORC interpretation will be indicated with a resumptive object pronoun clitic on the relative clause verb (Figure 1.2). The majority of the experiments reported here use stimuli where the matrix clause noun phrase and the relative clause noun phrase match in grammatical number and gender. Crucially, the only difference between an SRC and ORC in this format is the presence of the object resumptive pronoun clitic. However, the final experiment reported here uses stimuli where the matrix and relative clause noun phrases do not match grammatically; in these cases, both the inflection of the relative clause verb and the presence of a resumptive pronoun clitic differ between the two RC types.

It is important to note that these various strategies for relativization in MSA introduce a measure of ambiguity into the interpretation of relative clauses. When reading an SRC in MSA, a reader will read the relative clause verb followed by a noun phrase that is highly likely to be interpreted as the object of the SRC (Figure 1.2a). However, it is also plausible that the reader could interpret the relative clause noun phrase as the subject of a relative clause with VSO word order and assume a gapping strategy to mark the object rather than a resumption strategy (Figure 1.3). This means that, in sentences where the matrix and relative clause noun phrases are matched in grammatical number and gender and both grammatically

[1] Previous literature on RP processing in languages other than MSA has shown inconclusive evidence as to whether RPs help or hinder processing and comprehension (Meltzer-Asscher, 2021); however, RPs in MSA are grammaticalized and provide syntactic information (e.g., number and gender marking) that can aid in processing.
[2] While resumption is not required in every dialect of Arabic, previous research on the processing of resumptive pronouns in MSA has shown that a speaker's colloquial dialect has no significant effect on their ability to process resumption in MSA (Alresaini, 2016).

agree with the relative clause verb, SRCs are technically globally ambiguous with ORCs with VSO word order that utilize the gapping strategy.

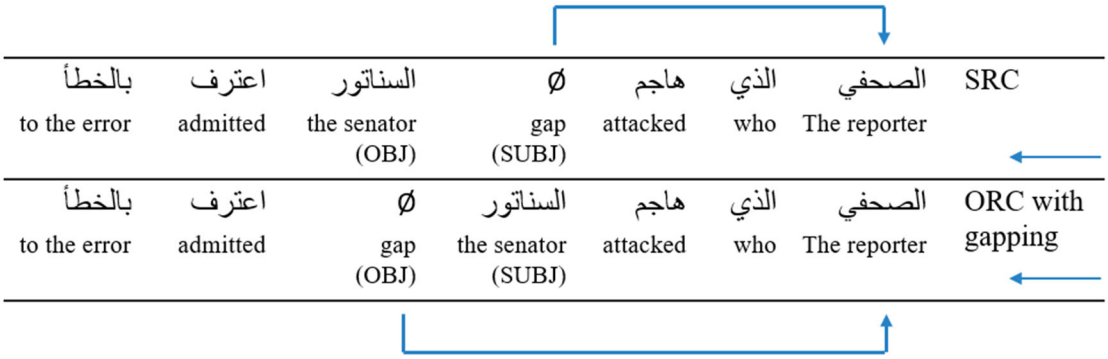

Figure 1.3: Sample SRC and ORC using the gapping strategy. Dependencies from the matrix clause subject to its gapped position in the relative clause are marked in blue.

There is also temporary ambiguity in ORCs that utilize the resumption strategy. MSA allows null object constructions in which a resumptive object pronoun clitic is used in lieu of an object noun phrase (e.g., "The reporter who attacked him…" where the null object for "him" must be resolved by context; see Figure 1.4) (Al-Sharafi & Gubaily, 2023). In this case, readers could read a relative clause verb with a resumptive object pronoun clitic and temporarily assume an SRC interpretation with a null object. This again is only a possible interpretation in sentences where the matrix and relative clause noun phrases match grammatically, and crucially, this interpretation becomes void upon reading the relative clause noun phrase after the relative clause verb. These ambiguities are considerations we take into account in our interpretation of findings for the first three experiments, and our final experiment contains a majority of stimuli that do not have sentences with grammatically matching matrix and relative clause noun phrases to help control for these ambiguities. We also explore the relative frequency of these types of structures in our corpus analysis in Section 2.1.1.

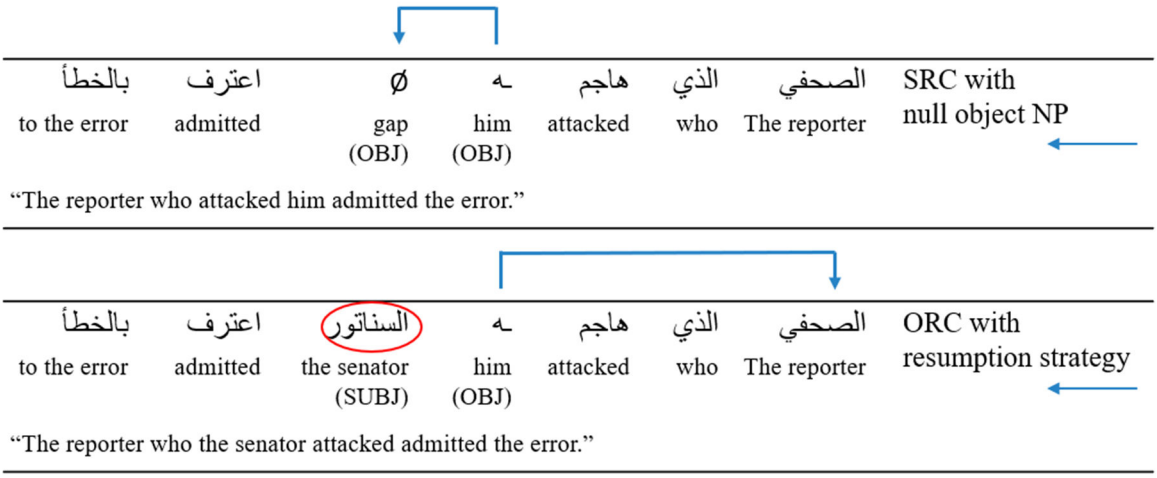

Figure 1.4: Sample SRC with an object pronoun and null object NP construction versus an ORC using the resumption strategy. Dependencies from the object pronoun clitic to its referent in each structure are marked in blue. An ORC using resumption is locally ambiguous with an SRC with a null object NP until encountering the relative clause NP, circled in red. At this point, the SRC interpretation is voided.

1.5. Dissertation studies

We first tested whether memory- or expectation-based theories best explained patterns of processing difficulty while reading relative clauses in Arabic (Chapter 2). We asked native Arabic speakers to read a series of SRCs and ORCs using a self-paced reading paradigm, and had them answer comprehension questions about what they read after reading each sentence. Our results showed that ORCs had longer overall reading times compared to SRCs, supporting predictions from expectation-based theories; however, the locus of processing difficulty was at the relative clause noun phrase, not the relative clause verb. Further, we analyzed the answers to the comprehension questions and found that participants were significantly more likely to incorrectly interpret an ORC than an SRC. We hypothesized that participants may be misreading ORCs as SRCs by skipping the resumptive pronoun clitic on the relative clause verb, or engaging in good-enough or noisy-channel processing by accepting a higher-frequency, and thus preferred, SRC interpretation.

As a first step toward exploring this question, we conducted a recall task using the same stimuli from the self-paced reading experiment (Chapter 3). We asked native Arabic speakers to read these sentences

and re-write each sentence word-for-word on a different page. If readers were misreading, we expected to see unidirectional errors of ORCs being mistakenly re-written as SRCs. However, if readers were instead engaging in good-enough or noisy-channel processing, we expected to see bidirectional errors of ORCs being mistakenly re-written as SRCs (due to SRCs being the more frequent structure), *and* SRCs being mistakenly re-written as ORCs (perhaps due to individual variance in semantic plausibility, or that SRCs are technically ambiguous with ORCs). We found that ORCs were misremembered as SRCs and SRCs were misremembered as ORCs, lending tentative support to theories of good-enough and noisy-channel processing.

We then conducted our first eye tracking experiment, which provides much more granular data behind processing behaviors than self-paced reading, to answer our two outstanding questions (Chapter 4). First, we wanted to further explore the locus of processing difficulty in Arabic relative clauses and determine whether Arabic readers paid the cost for updating their expectations at the relative clause verb or at the relative clause noun phrase. Second, we wanted to explicitly test whether readers were misreading ORCs, or simply accepting a preferred SRC interpretation. We again had participants read various SRCs and ORCs in Arabic and answer comprehension questions after each sentence, and used a combination of eye movement behaviors and answers to comprehension questions to diagnose misreading versus noisy-channel processing. Our findings showed that readers were indeed registering the resumptive object pronoun clitic at the relative clause verb, and thus were not misreading ORCs. Further, processing behaviors for trials in which ORCs were misinterpreted showed that readers did not pay the processing cost of updating their expectations as they did in trials where ORCs were accurately interpreted, providing further support for good-enough and noisy-channel processing. Finally, we found that this additional processing cost for ORCs is indeed paid at the relative clause noun phrase, and not at the relative clause verb, suggesting that Arabic readers are not engaging in the most incremental, probabilistic processing possible.

Finally, we investigated the relationship between statistical expectations (i.e., of a certain clause type) and grammatical cues (i.e., number and gender marking) to identify whether increasing the number of cues in favor of a veridical interpretation could overcome the statistical expectations for a higher-frequency structure (Chapter 5). We conducted a second eye-tracking experiment and created new versions of our stimuli where the matrix and relative clause noun phrases either matched grammatically, were mismatched on number, or were mismatched on both number and gender, and analyzed whether sentences with grammatically matched noun phrases had more noisy interpretations than sentences with two grammatical mismatches, and thus more cues in favor of a veridical ORC interpretation. We found that the number of grammatical cues in favor of a veridical ORC interpretation did not have a significant effect on the number of noisy interpretations of ORCs, nor on the processing behaviors while reading these sentences. Thus, grammatical cues were not strong enough to overcome statistical expectations in a good-enough or noisy-channel processing framework.

Chapter 2

Experiment 1: Self-paced reading

2.1. Introduction

We first tested whether memory- or expectation-based theories of sentence processing best explained reading patterns in Arabic SRCs and ORCs through a self-paced reading task. Memory- and expectation-based theories make different predictions about which structure should be harder to read in Arabic, so we compared average reading times for each structure to determine which was harder to process and how this aligned with predictions from each theory. Native Arabic speakers were presented with various SRCs and ORCs using a self-paced word-by-word reading paradigm. Reading or reaction times (RTs) were collected for each word and used as an indication of processing difficulty.

Figure 2.1: Arabic matrix clause subject dependency in the (a) SRC and (b) ORC condition. The probabilistic disambiguating region, the relative clause verb, is circled in red. The red vertical line on the ORC verb delineates the resumptive pronoun clitic from the relative clause verb.

2.1.1. Corpus analyses

Prior to beginning our experiments, we conducted a number of corpus analyses to explore the frequency of the various forms of relative clauses that occur in Arabic. We used the Penn Arabic Treebank Part 3 v 3.2 corpus (Maamouri et al., 2010) for all our analyses. The corpus was created using newswire data from the Lebanese news agency *An Nahar* and contains over 402K tokens that are annotated for POS and syntactic treebanking.

We first investigated whether SRCs or ORCs were more common in MSA in order to determine what expectation-based theories would predict would be the easier structure to process. We used two search parameters. We first searched for relative clauses, labeled with the SBAR tag, with an explicit "who" relative pronoun (الذي plus all other gender and number declensions), labeled with the WHNP and REL_PRON tag. Each WHNP tag contained a numeral trace that matched the NP to which it belonged (e.g., WHNP-2 would trace back to NP-2). We also used a second, broader search parameter that included relative clauses that used either an explicit "who" relative pronoun (SBAR plus WHNP REL_PRON), no relative pronoun (SBAR plus WHNP -NONE-), or the more general "that" (ما) relative pronoun (SBAR-NOM plus WHNP). We identified 2,928 relative clauses using our first search parameter, of which 71% were SRCs. When using our broader second search parameter, we identified 7,268 relative clauses, of which 79% were SRCs. We thus concluded that SRCs are much more common than ORCs in MSA.

We then investigated whether relative clause word order was more likely to be VSO or SVO to inform the design of our materials. While both word orders are grammatical within a relative clause, our native speaker consultants stated that they highly preferred the VSO word order. (NB: relative clause word order in SRCs would appear to be the same on the surface for both VSO and SVO word order, but this difference would become apparent in ORCs). To do this, we searched all the relative clauses identified in the previous analysis to see how often the subject (NP-SUBJ) occurred before the verb (VP) within the relative clause. Our results showed that 98% of relative clauses were verb-initial, confirming our native speakers' intuitions about VSO word order.

Finally, we wanted to investigate the frequency of relative clauses in MSA that are temporarily ambiguous at the relative clause verb. Section 1.4.2 discussed how encountering a relative clause verb with a resumptive object pronoun clitic may signal an ORC, but it could also signal an SRC with a null object construction (when the matrix and relative clause nouns are grammatically matched). We thus wanted to determine how often this null object construction appeared among the SRCs in our dataset. Following our initial corpus analysis, we first used a search parameter to identify relative clauses with an explicit "who" pronoun, then a second parameter to broadly include relative clauses that used either an explicit "who" relative pronoun, no relative pronoun, or the more general "that" relative pronoun. We then searched within our identified SRCs to see if any of them included a direct object pronoun clitic (IVSUFF_DO) that was labeled as the object within the clause (NP-OBJ). We manually reviewed the results of this search to see if any of the identified clauses had this resumptive object pronoun clitic along with a trace parameter for a null object (NP-# (-NONE- *T*) where # represents the numeric trace to a NP that *does not* match the subject NP trace). We found one example of this type of structure within our first search parameter (out of 2,073 total SRCs), and three more examples within our broader, second search parameter (out of 5,725 total SRCs). While we do find evidence of this construction occurring in MSA, it appears to be much less frequent than the standard SRC construction with an explicit object NP.

Based on the data from this corpus, reading a relative clause verb without a resumptive pronoun should strongly suggest an SRC interpretation while reading a relative clause verb with a resumptive pronoun should strongly suggest an ORC interpretation; however, there is a possibility that readers will entertain other possible, but less probable based on the data, interpretations. We consider the relative clause verb region to be the probabilistic, but not deterministic, disambiguating region in our theoretical predictions based on these findings.

2.1.2. Theoretical predictions

Memory-based theories grounded in the DLT predict comparable processing times for SRCs and ORCs in Arabic. This is due to the inclusion of the resumptive object pronoun clitic on the relative clause verb, the

probabilistic disambiguating region, in the ORC condition: because the resumptive pronoun is cliticized to the relative clause verb, the dependency between the disambiguating region and the matrix clause subject have the same total length in both clause conditions (Figure 2.1). The locus of processing difficulty – i.e., where the reader pays the integration cost – should thus probabilistically be at the relative clause verb.

On the other hand, expectation-based theories predict faster processing times in SRCs as it is the more frequent structure in MSA, as we confirmed through our corpus analysis. The locus of processing difficulty – i.e., where the reader pays the cost of updating their expectations for an ORC interpretation – should also probabilistically be at the relative clause verb, as the presence of a resumptive pronoun clitic should strongly suggest an ORC interpretation.

2.2. Methods

2.2.1. Participants

Forty-eight native Arabic speakers proficient in MSA (20 women, 26 men, 2 not reported; mean age: 27; *SD* = 7.25) were recruited from Prolific[3]. Participants were paid $4.50 for their participation. Prior to beginning the experiment, participants completed a questionnaire which included an obligatory question about their native language and optional questions about their demographic backgrounds (see Appendix A for the full list of questions). Participants were considered native Arabic speakers if they selected either MSA or another dialect of Arabic as their native language. Two participants were excluded for not matching this criterion. Of the remaining participants, the participant pool represented a diverse sample of Arabic dialects: 14 different dialects were reported across the 44 participants who completed the voluntary portion of the questionnaire. We also established an a priori criterion to exclude any participant who scored lower than 75% accuracy on sentence comprehension and filler comprehension questions

[3] https://www.prolific.com/

(i.e., not questions that targeted relative clause comprehension), which resulted in the exclusion of one participant.

2.2.2. Materials

The stimuli were designed to investigate processing behaviors in VSO word order and with grammaticalized resumptive pronouns, and utilized the flexible word order in Arabic to minimize variation between SRCs and ORCs. For each sentence, the matrix clause was SVO and the relative clause was VSO. This word order was selected with the help of native speaker input and confirmed through a corpus analysis. Given these word orders, readers first read the matrix clause subject, followed by the relative pronoun, and then the relative clause verb in both conditions. The key difference between the SRC and the ORC condition was the presence of the resumptive object pronoun in the ORC condition as a bound clitic on the relative clause verb (Figure 2.1).

Stimuli were adapted and translated from previous studies on relative clause processing (Gordon et al., 2001; Staub, 2010; Traxler et al., 2002). Arabic nouns, verbs, and pronouns are marked for both number and gender, so matrix and relative clause nouns were matched on number and gender so that the subject of the relative clause would not be disambiguated by number and gender marking on the relative clause verb. Nouns were either masculine singular or feminine singular, and gender was balanced across items. In ORC items, the resumptive object pronoun for masculine singular nouns was one character (ـه) while the resumptive object pronoun for feminine singular nouns was two characters (ـها). All nouns were definite and animate to control for animacy effects (Mak et al., 2002; Traxler et al., 2002; 2005). Finally, all stimuli were presented in a non-diacritized format, as is standard for written publications in MSA (Hermena et al., 2015). Individual words were diacritized to avoid ambiguity when necessary.

A norming study was conducted to confirm that the subject and object of each relative clause were equally plausible in both clause conditions (e.g., “the reporter attacked the senator” is as plausible as “the senator attacked the reporter”). Native Arabic speakers ($N = 76$; 33 women, 42 men, 1 not reported; mean

age: 28; *SD* = 6.65) were recruited through Prolific and asked to rate the plausibility of 45 sentences on a Likert scale (1 = highly implausible, 7 = highly plausible). These participants were excluded from participating in the self-paced reading experiment, and therefore did not overlap with the self-paced reading experiment participants. Plausibility ratings were collected for both the full stimuli sentences (e.g., "The reporter who attacked the senator admitted the error") and the relative clauses as simplified transitive sentences (e.g., "The reporter attacked the senator"). The study also included an equal number of filler items, some of which were implausible distractor sentences that functioned as attention checks. Four stimuli were excluded after a paired t-test revealed substantial discrepancies between plausibility ratings in the SRC and ORC conditions for those items, and one stimulus was excluded for low overall ratings. These exclusions resulted in 40 total stimuli. The mean plausibility rating for full stimuli sentences for the remaining items was 6.11 (*SD* = 0.49) for SRCs and 6.00 (*SD* = 0.57) for ORCs, and the mean plausibility rating for the simplified transitive sentences was 6.18 (*SD* = 0.46) for SRCs and 6.08 (*SD* = 0.49) for ORCs.

In addition to the 40 target sentences, 80 unrelated filler sentences were included. Comprehension questions appeared after all 40 experimental sentences and 20 filler sentences (i.e., half of all items). Of the 40 experimental stimuli comprehension questions, half of the questions targeted comprehension of the relative clause (e.g., "Did the reporter attack the senator?" (a) "Yes," (b) "No"), and the other half targeted comprehension of the sentence overall (e.g., "Did the reporter admit the error?" (a) "Yes," (b) "No"). Whether "Yes" or "No" was the correct answer was balanced within question type. In total, each participant read 120 sentences (40 target sentences (20 for each clause type) + 80 filler sentences) and answered 60 comprehension questions. Experimental items were counterbalanced in a Latin square design, and the order of sentences was randomized separately for each participant.

2.2.3. Procedure

The study was conducted online through Ibex Farm[4], a site for hosting psycholinguistic experiments. Participants were told that they would be reading sentences in Arabic and answering comprehension questions. All experimental instructions were given in MSA. Prior to the start of the experiment, participants saw two practice stimuli and answered one practice comprehension question. Each sentence was presented word-by-word using a subject-paced paradigm in which participants used the spacebar to advance through the sentence. Each word in the sentence was presented in isolation with no option to move backward in the sentence. RTs were collected at each key stroke indicating the appearance of the next word in the sentence. The experiment took about 30 minutes on average.

2.3. Analysis and Results

Prior to our analysis, RTs faster than 100 ms and slower than 2,000 ms were excluded. This resulted in a total data loss of 4.6%. RTs were then residualized within-subject to control for word length effects (Ferreira and Clifton, 1986). Outliers outside of 3 standard deviations from the mean were also excluded, resulting in an additional 2.1% data loss. Each sentence was divided into target regions for our analysis (Figure 2.2b). These regions included the matrix noun phrase through the matrix verb, plus one to three spillover regions (one word each) depending on the length of the sentence. The key regions of interest in our analysis were Region 2 (the relative clause verb, plus the resumptive pronoun clitic in the ORC condition), Region 3 (the relative clause noun phrase), and Region 4 (the matrix clause verb). Region 2 is where the reader should be able to probabilistically disambiguate between an SRC and an ORC reading, and so is the locus of processing predictions from memory- and expectation-based theories. We further focus on Regions 3 and 4 to capture any potential spillover effects.

Average residualized RTs in each region by clause type are plotted in Figure 2.2a. Raw RTs, trending in similar directions as the residualized RTs, are additionally given in Appendix C. Negative residual

[4] https://adrummond.net/ibexfarm

RTs indicate shorter processing times given word length, and positive residual RTs indicate longer processing times given word length.

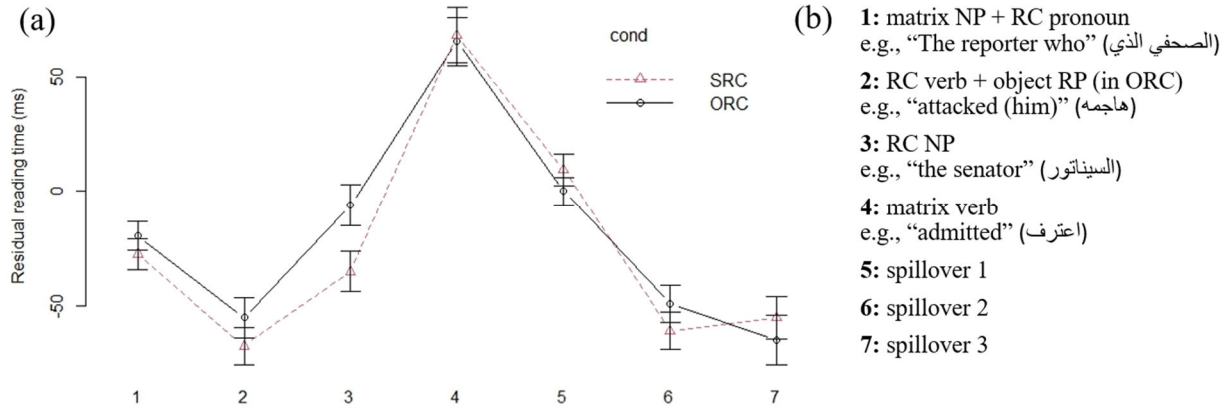

Figure 2.2: (a) Average residualized RTs for each region by clause type (after data preprocessing); (b) Regions of interest with Arabic examples and their English gloss.

To determine whether SRCs or ORCs were more difficult to process, we first summed RTs for each trial across our region of interest (Regions 2-4) and compared RTs by clause type. We then analyzed RTs for each individual region by clause type. For each analysis, we first looked at outcomes from all RT data, and then at outcomes from RT data for items with correct comprehension question answers.

Data were analyzed using the *brms* package in R (Bürkner, 2017). Linear mixed-effects regression models were fit for within-subject residualized RTs in Regions 2-4, plus each individual region. The models included a sum-coded fixed effect for Clause Type (ORC: 1, SRC: -1). We used the maximal random effects structure by subjects and items as justified by the design (Barr et al., 2013), resulting in random intercepts for Participant and Item, and random slopes by Clause Type for both Participant and Item. We consider the model estimates as significant if the credible interval (CrI) does not include 0, or over 95% of the sampled posterior distribution is over or under 0 in the predicted direction. Model estimates for main effects are reported in Table 2.1.

Table 2.1: Linear mixed-effects model estimates of the dependent variable, Clause Type, on residualized RTs by region, including SE estimates and CrIs. The % > 0 column shows the percent of the sampled posterior distribution that is over 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates marked with an asterisk are significant.

	All Items				Correct Items			
Region	**β**	***SE***	**CrI**	**% > 0**	**β**	***SE***	**CrI**	**% > 0**
Regions 2-4	21.81*	13.24	[-4.14, 46.98]	95	22.90	14.71	[-6.19, 51.83]	94
Region 1	7.90	7.67	[-7.62, 22.82]	86	6.64	8.65	[-10.36, 23.64]	78
Region 2	5.80	5.56	[-5.04, 16.82]	86	5.83	6.23	[-6.14, 18.17]	83
Region 3	17.63*	7.15	[3.49, 31.71]	99	23.42*	8.50	[7.05, 40.51]	100
Region 4	-0.07	8.16	[-16.03, 15.93]	49	-4.36	9.47	[-23.4, 14.13]	32
Region 5	-4.47	5.76	[-16.00, 7.08]	21	-4.00	6.60	[-16.83, 8.99]	27
Region 6	5.15	5.85	[-6.40, 16.30]	81	10.16	6.82	[-3.34, 23.61]	93
Region 7	-3.55	6.23	[-15.78, 8.96]	29	-0.45	7.25	[-14.87, 14.02]	48

Model estimates for the summed Regions 2-4 showed a significant effect by Clause Type on RTs, such that ORCs were read significantly longer than SRCs on average in this region. This effect was significant across all items, but was just nearly significant for only correct items (94% of the sampled posterior distribution greater than 0). The individual region models showed that this effect was also individually significant in Region 3 (the relative clause noun), and this was consistent across all items and for only correct items. Notably, there were no significant effects for Clause Type in Region 2 (the relative clause verb) for all items or only correct items.

2.3.1. Re-investigating the relative clause verb

Despite finding no significant effects for Clause Type at the relative clause verb, we wanted to see whether readers were paying any sort of additional processing cost for reading the resumptive pronoun clitic on the relative clause verb in ORCs. Our analyses used residualized reading times which controls for word length, so it was possible that this extra processing cost was present but proportional to length. To investigate this possibility, we re-ran our models in Region 2 (the relative clause verb) using the raw reading times instead of the residualized reading times. Model estimates are reported in Table 2.2.

Table 2.2: Linear mixed-effects model estimates of the dependent variable, Clause Type, on raw RTs in Region 2, the relative clause verb, including SE estimates and CrIs. Estimates marked with an asterisk are significant.

	All Items			Correct Items		
Region	**β**	***SE***	**CrI**	**β**	***SE***	**CrI**
Region 2	23.83*	6.00	[12.27, 35.64]	24.36*	6.82	[11.20, 38.16]

Model estimates showed significant effects by Clause Type on raw RTs at the relative clause verb, such that ORCs were read significantly longer than SRCs on average. This effect was significant both for all items and for only correct items. This suggests that readers do pay an additional processing cost for reading the relative clause verb in ORCs; however, this additional cost is proportional to reading the extra character(s) from the resumptive object pronoun clitic.

2.3.2. Participant accuracy on comprehension questions

Our initial review of comprehension question accuracy revealed higher-than-expected error rates. We observed that participants performed markedly worse on comprehension questions that targeted relative clause comprehension (i.e., the correct SRC or ORC interpretation; 75.4% correct after 3 questions were excluded; see below) than on questions that targeted overall sentence comprehension (93.2% correct). Participants were also much more likely to get the correct interpretation of an SRC (83.3% correct; 349 trials) than an ORC (67.4% correct; 279 trials). It is important to note that because resumptive object pronouns are not strictly required in these structures, the SRC interpretation is technically ambiguous and could be read as ORCs; so, while the comprehension question accuracy rates show that people overwhelmingly interpret them as SRCs, the "incorrect" answers are not technically wrong. However, the ORCs are *only* globally interpretable as ORCs, and yet have even lower comprehension accuracy. To investigate the discrepancy in comprehension question accuracy for SRCs and ORCs, we analyzed comprehension question answers by clause type (SRC vs. ORC) and question type ("Yes" correct answer vs. "No" correct answer).

Forty comprehension questions were used for the stimuli in the experiment: 20 that targeted relative clause comprehension and 20 that targeted overall sentence comprehension. For our analysis, we focused only on questions that targeted relative clause comprehension. Of those 20 questions, 3 questions were excluded from further analysis: two were excluded due to experimental coding issues and one was excluded for issues related to poor translation from English, creating an awkward sentence in MSA.

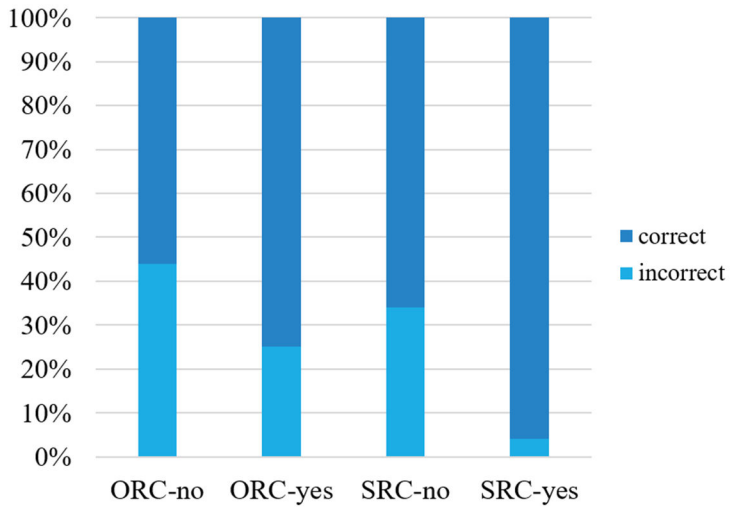

Figure 2.3: Proportion of correct answers by clause type (SRC or ORC) and correct answer condition (correct answer is "yes" or "no").

We first investigated the proportion of correct comprehension question answers by clause type and correct answer condition (Figure 2.3). Whereas the "no" condition was relatively comparable across clause type, there appeared to be a substantial discrepancy between clause types for the "yes" condition. Participants appeared to suffer from a "yes" bias when answering comprehension questions about SRCs – SRCs with a "yes" comprehension question had significantly fewer incorrect answers – but this difference was not as marked with ORCs. This would entail that after reading an ORC stimulus (e.g., "The reporter who the senator attacked admitted the error."), if prompted with the veridical ORC interpretation by the comprehension question (e.g., "Did the senator attack the reporter?"), a reader was still fairly likely to reject the veridical interpretation and answer "no" when the correct answer was "yes."

To evaluate the statistical significance of these findings, a logistic mixed-effects regression model was fit to the data with Correctness as the dependent variable and Clause Type (ORC: 1, SRC: -1) and Correct Answer (No: 1, Yes: -1) as sum-coded fixed effects, plus their interaction. We also used the maximal random effects structure by Participant and Item. Model estimates are reported in Table 2.3.

Table 2.3: Logistic mixed-effects model estimates of each dependent variable and interaction on correct comprehension question answers, including SE estimates and CrIs. Estimates marked with an asterisk are significant.

Main Effects	**β**	***SE***	**CrI**
Clause Type	-0.82*	0.19	[-1.26, -0.49]
Correct Answer	-0.98*	0.20	[-1.40, -0.62]
*Clause Type * Correct Answer*	0.56*	0.19	[0.22, 0.98]

Model estimates showed a significant effect of both Clause Type and Correct Answer on Correctness. First, comprehension questions for ORC stimuli had significantly more incorrect answers. Additionally, questions whose correct answer was "no" received significantly more incorrect answers. There was an additional significant effect for the interaction between Clause Type and Correct Answer, resulting in a subadditive effect for our predictor variables.

2.4. Discussion

We set out to test predictions from memory- and expectation-based theories of sentence processing in Modern Standard Arabic, using subject- and object-extracted relative clauses as a test case. We considered the disambiguating region, and thus the locus of processing difficulty, to be the relative clause verb, as our corpus results showed that seeing a resumptive object pronoun clitic on the relative clause verb should strongly suggest an ORC interpretation. Based on this consideration, memory-based theories predicted comparable processing times due to the use of resumptive object pronoun clitics in Arabic: given the inclusion of these clitics, the dependency from the matrix clause subject to the relative clause can be resolved upon reading the relative clause verb, regardless of whether the matrix clause noun is the

subject or the object of the relative clause. On the other hand, expectation-based theories predict longer processing times for ORCs as they are less frequent than SRCs in Arabic.

We found that ORCs are read more slowly than SRCs in Arabic. This effect was significant when averaged across the relative clause verb, noun, and matrix clause verb, and was also individually significant at the relative clause noun. These results support predictions from expectation-based theories of sentence processing; however, the locus of processing difficulty was not where it was predicted to be based on our corpus analyses. The relative clause verb was the first region that allowed the reader to probabilistically disambiguate between an SRC and an ORC, and thus was where we expected participants to pay the additional processing cost of updating their expectations when reading an ORC. Rather, our data showed that this processing cost was paid at the relative clause noun, the region directly following the verb. One possible explanation for this phenomenon is that our data showed processing delays that are common in self-paced reading tasks (Just et al., 1982; Frank et al., 2013), where the reading time difference in a critical region carries over into the following word or region. It is possible, then, that participants may experience processing difficulty at the relative clause verb, but these reading time differences instead appeared at the relative clause noun. However, it is also possible that participants are not probabilistically disambiguating between an SRC and an ORC at the relative clause verb, as predicted by expectation-based theories that model strict incremental processing. Our subsequent experiments utilized eye-tracking in place of self-paced reading in order to minimize any possible spillover effects, and further explore the question of where Arabic readers experience processing difficulty when reading ORCs.

While our results show strong support for expectation-based processing costs, it is important to note that these findings do not necessarily preclude the possibility of memory-based processing costs in language processing in Arabic. Given our stimuli design, memory-based theories predicted comparable processing costs for both SRCs and ORCs, and so our results do not rule out the possibility of memory constraints contributing to processing difficulty in structures with different dependency lengths. Future

research could utilize different syntactic structures with varying dependency lengths to investigate the possible contributions of memory limitations to language processing in Arabic.

We additionally analyzed our comprehension question data to investigate patterns of language comprehension in our stimuli. We were particularly interested in whether participants had a harder time understanding ORCs compared to SRCs. Our study included 20 questions that targeted relative clause comprehension (i.e., whether "the reporter attacked the senator" or "the senator attacked the reporter"), and each question had simple "yes" or "no" answer options, in line with what is commonly used in noisy-channel processing literature (Huang & Staub, 2021). Our analysis showed that ORCs received significantly more incorrect comprehension question answers than SRCs, and thus were misinterpreted at substantially higher rates than SRCs. Further, we found that questions whose correct answer was "no" also received significantly more incorrect answers, suggesting that participants were engaging in a "yes" bias and were more likely to respond "yes" than "no," regardless of the correct answer. Finally, we found an additional subadditive effect for clause type and correct answer condition. While our data showed that SRCs were especially susceptible to the "yes" bias, it appeared that ORCs were not as sensitive to this bias: SRCs with "yes" answers had 96% accuracy compared to ORCs with "yes" answers at only 75% accuracy. We do not have a specific explanation for this subadditive effect, and it is important to note that the findings are based on a relatively small sample size of only 17 questions. However, taken together, the data indicate that readers frequently misinterpret these sentences, and the framing of the comprehension questions appears to bias the results. Future research on noisy-channel processing should consider how the framing of comprehension questions can bias noisy interpretations, as it appears that the "yes"/"no" answer paradigm is accepted practice in the literature (e.g., Cutter et al., 2024; Gibson et al., 2013; Ryskin et al., 2018).

Given these data, we hypothesized that some readers mistakenly interpret ORCs as SRCs while reading. We considered two possible reasons for this misinterpretation. First, it is possible that the resumptive object pronoun clitic on the ORC verb is short enough that it is missed during reading. Our

native speaker consultants all observed that the resumptive pronoun clitic is easy to miss, and missing this clitic would effectively result in reading an SRC. On the other hand, it is also possible that readers register and read the resumptive pronoun clitic, but reject an ORC interpretation in favor of a higher-frequency SRC interpretation. This hypothesis is in line with theories of good-enough and noisy-channel processing (Ferreira et al., 2002; Levy, 2008a), which state that readers will sometimes accept a more expected or higher-probability sentence structure over a less expected one. Recent research in Hebrew relative clauses – a language that is typologically similar to Arabic – demonstrated that readers prefer high-frequency but grammatically incorrect interpretations of sentences to their grammatical but infrequent counterparts, suggesting that expectations strongly modulate processing (Keshev & Meltzer-Asscher, 2021). Further, previous reading studies have found that re-reading does not improve comprehension accuracy (Christianson et al., 2017); thus, even when given the opportunity to re-read a relative clause verb with a resumptive pronoun clitic, readers may not update their understanding of the sentence. These findings lead us to suspect that readers are engaging in good-enough or noisy-channel processing while reading ORCs in Arabic.

Chapter 3

Experiment 2: Recall task

3.1. Introduction

We took first steps toward investigating whether misinterpretations were due to misreading or good-enough/noisy-channel processing by conducting a recall task (Bock & Brewer, 1974; Flores D'Arcais, 1974; James et al., 1973). Using the same stimuli as our self-paced reading task, we asked participants to read each sentence and then reproduce the sentence word-for-word from memory. If readers are misreading ORCs due to missing the resumptive pronoun clitic on the relative clause verb, we expect to see unidirectional errors of ORCs reproduced as SRCs; missing the resumptive pronoun would simply result in an SRC reading and interpretation. However, if readers are correctly reading ORCs but accepting noisy SRC interpretations, we expect to see both ORCs misremembered as SRCs and SRCs misremembered as ORCs; ORCs may be noisily interpreted as SRCs as they are the more frequent structure, but SRCs may also be noisily interpreted as ORCs if semantic expectations outweigh syntactic expectations (Gibson et al., 2013). Our sentences were normed so that the SRC and ORC interpretations of each sentence should generally be equally plausible; however, participants may have idiosyncratic preferences about the plausibility of a given interpretation.

3.2. Methods

3.2.1. Participants

Eighty native Arabic speakers proficient in MSA (36 women, 43 men, 1 genderfluid; mean age: 29; *SD* = 7.32) were recruited from Prolific. Users who participated in the self-paced reading task or the norming task were not eligible to participate in the recall task. Participants were paid $4 for their participation. Prior to beginning the experiment, participants completed the same questionnaire as was included in the self-paced reading task. We established the same a priori criteria to exclude participants based on native language and comprehension question accuracy, but no participants fit these criteria.

3.2.2. Materials

We used the same stimuli from the self-paced reading task for the recall task. Each participant saw 40 stimuli (20 SRCs + 20 ORCs) and 40 filler items for a total of 80 items. Experimental items were split into four lists and counterbalanced within and across lists in a Latin square design. Participants were randomly assigned to one of four lists, and the order of sentences within each list was randomized separately for each participant.

3.2.3. Procedure

The study was conducted online through Qualtrics. Participants were told that they would be reading sentences in Arabic and rewriting them from memory. All experimental instructions were given in MSA. Prior to the start of the experiment, participants completed one practice trial. For each block, one sentence was presented in its entirety for a maximum of 10 seconds before auto-advancing to the next page. The auto advance feature was enabled to discourage any screen-shotting or copying of the sentence for the task. Participants had the option to continue to the next page at any time before the 10 seconds lapsed. On the next page, the participant was asked to rewrite the sentence word-for-word from memory into a text box. Participants completed 80 blocks, one for each experimental item. The experiment took 45 minutes on average.

3.3. Analysis and Results

Data were manually reviewed and hand coded for correct replications. Replications were coded as being correct if the correct clause type was reproduced in the form that it was written in the sentence, regardless of any other errors in the sentence, or incorrect if the incorrect clause type was reproduced. Two stimuli were excluded due to errors with presentation within the experiment. Any trial in which the participant did not re-write a complete sentence was also excluded. This resulted in a total data loss of 10%.

Overall, misremembrance of the correct clause type was low (<4% incorrect items). However, within those items, there was a sizable difference in misremembrance by clause type. 71% of errors (33 trials) were ORCs being misremembered as SRCs, and the remaining 29% of errors (13 trials) were SRCs being misremembered as ORCs. Participants did very well on this task, and as such these findings were derived from a fairly small sample size; however, the data demonstrated that ORCs were much more likely to be misremembered than SRCs, and clause type errors were made in both directions (i.e., both SRCs and ORCs were misremembered as opposed to just ORCs).

To evaluate the statistical significance of these findings, a logistic mixed-effects regression model was fit to the data with Correctness as the dependent variable, and Clause Type (ORC: 1, SRC: -1) and Verb Length (numeric, centered) as sum-coded fixed effects, plus their interaction. Verb Length was the length of the verb not including the resumptive pronoun clitic, so this metric was consistent across clause types. We also used the maximal random effects structure by Participant and Item. Model estimates for main effects are reported in Table 3.1.

Table 3.1: Logistic mixed-effects model estimates of each dependent variable and interaction on recall task correctness, including SE estimates and CrIs. Estimates marked with an asterisk are significant.

Main Effects	**β**	***SE***	**CrI**
Clause Type	-1.70*	0.74	[-3.51, -0.61]
Verb Length	-0.44	0.38	[-1.24, 0.26]
*Clause Type * Verb Length*	0.18	0.37	[-0.54, 0.90]

Model estimates showed a significant effect for Clause Type on successful recall rates, such that ORCs were significantly less likely to be remembered correctly than SRCs. Neither Verb Length nor the interaction between Clause Type and Verb Length were significant.

3.4. Discussion

This recall task took first steps toward investigating the systematic misinterpretation of ORCs that we observed in our self-paced reading task. We hypothesized that misinterpretations could be due to the reader misreading the relative clause verb and missing the resumptive object pronoun clitic, which should probabilistically signal an ORC interpretation, or due to the reader accepting a preferred SRC interpretation despite registering the resumptive pronoun clitic on the relative clause verb. Our recall task tested whether only ORCs were misinterpreted as SRCs, suggesting that readers may simply be misreading ORCs, or whether both clause types were misremembered, suggesting that expectations or other processing mechanisms may be at play.

The outcomes from our experiment showed that few items were misremembered overall, but when they were, ORCs were misremembered as SRCs *and* SRCs were misremembered as ORCs. Further, ORCs were significantly more likely to be misremembered as SRCs than the other way around. We interpreted these findings as potential support for good-enough or noisy-channel processing over misreading. Whereas misreading would result in unidirectional misinterpretations of ORCs as SRCs, both clause types were misremembered as the other type, suggesting that the misinterpretations of ORCs are not accidental misreadings.

Chapter 4

Experiment 3: Eye tracking

4.1. Introduction

Our next study built upon our previous self-paced reading task and recall task by explicitly investigating what causes ORC misinterpretations. We did this by conducting an eye tracking experiment, which provides more granular measures of processing behavior than self-paced reading. We made three changes to our design to better examine these misinterpretations. First, we substantially increased the number of experimental items that the participants read. Since our primary analyses will focus on the subset of items that were misinterpreted, we increased the total number of items to also increase the number of misinterpreted items proportionally. Second, we included a comprehension question after every experimental item that specifically tested the reader's interpretation of the relative clause. Our first experiment used a combination of questions targeting relative clause interpretation and overall sentence interpretation, and this precluded us from investigating relative clause interpretations for half of our stimuli. Finally, we changed the response options for the comprehension questions from "Yes" or "No" to be less biased toward a positive or negative response. The results from our first experiment showed that the framing of the comprehension questions biased the resulting interpretations, so we controlled for this bias by using a different question format.

The goals of this experiment were twofold. We first revisited our initial research question and assessed whether ORCs were harder to process than SRCs, in line with our previous findings and in support of expectation-based theories, and whether we observe this difference in processing difficulty at the relative clause verb or the relative clause NP. We did this by comparing processing behaviors in SRCs versus ORCs in items with correct comprehension question answers, as items with correct comprehension questions reflect veridical processing behaviors. We also considered which distinct eye movement measures reflect processing difficulty due to expectations, following Staub (2010).

We then asked whether readers were initially misreading ORCs as SRCs by missing the resumptive object pronoun clitic, or correctly reading ORCs yet accepting noisy SRC interpretations (Keshev & Meltzer-Asscher, 2021). To answer this question, we focused on processing behaviors in ORCs with incorrect comprehension question answers, and investigated whether these processing behaviors were more similar to behaviors for correctly interpreted ORC items or correctly interpreted SRC items. Misreading would suggest that misinterpreted ORCs would be read like correct SRCs, as misreading and missing the resumptive pronoun would result in reading an SRC (because the resumptive pronoun is the only difference between the two). On the other hand, good-enough and noisy-channel processing theories would suggest that misinterpreted ORCs would be read like correct ORCs, but be misinterpreted due to competing expectations.

Table 4.1: Predicted behavioral outcomes for veridical processing, misreading, and good-enough/noisy-channel processing.

Theory	**Comparison**	**Predicted behavior at disambiguating region (RC verb)**
Veridical processing	Cor ORC vs. Cor SRC	Increased processing difficulty
Misreading	Incor ORC vs. Cor ORC	Decreased processing difficulty due to missing RP clitic
	Incor ORC vs. Cor SRC	No significant difference; ORC misread as SRC
Good-enough/ noisy-channel processing	Incor ORC vs. Cor ORC	No significant difference; both read without missing RP clitic
	Incor ORC vs. Cor SRC	Increased processing difficulty; incorrect ORC read similarly to correct ORC

We conducted three analyses (Table 4.1). First, we compared SRC and ORC trials with correct comprehension question answers (Cor ORC vs. Cor SRC) to determine differences during veridical processing. Based on our understanding of the distributional statistics of SRCs and ORCs in Arabic from our corpus analysis, predictions from the processing theories investigated here all point to the relative clause verb as the primary region of interest. Memory-based theories predict equal processing difficulty between SRCs and ORCs at the relative clause verb: due to the inclusion of the resumptive object pronoun, the matrix noun can probabilistically be integrated upon seeing the verb regardless of whether it is the subject or object of the relative clause (Figure 2.1). On the other hand, expectation-based theories predict increased processing difficulty at the relative clause verb as this is where the clause can be probabilistically disambiguated between an SRC or ORC (Figure 1.2). If Staub's (2010) results on different processing behaviors from memory- versus expectation-based difficulties extend to Arabic, then we would expect any observed expectation-based processing difficulty to manifest in increased regressive saccades.

Our next two analyses investigated misreading versus good-enough/noisy-channel processing by comparing misinterpreted ORC trials to both correct ORC trials (Incor ORC vs. Cor ORC) and correct SRC trials (Incor ORC vs. Cor SRC). If readers are misreading the ORC verb by missing the resumptive pronoun, then in those cases the ORC verb would be read as identical to the SRC verb. We would then see no significant difference between misinterpreted ORCs and correct SRCs at the relative clause verb. However, we *would* see a significant difference between misinterpreted and correct ORC items, as only correct ORC items pay the additional processing cost of reading the resumptive pronoun. On the other hand, if good-enough or noisy-channel processing is occurring, then misinterpreted ORCs would be read similarly to correct ORCs at the relative clause verb, with readers paying the cost of reading the resumptive pronoun for all ORCs. Incorrect ORC interpretations would then arise from later good-enough or noisy-channel processing. In this case, we would see no significant difference between misinterpreted

and correct ORC trials at the relative clause verb, but a significant difference between misinterpreted ORC and correct SRC trials in the same region.

4.2. Methods

4.2.1. Participants

Forty-seven native Arabic speakers proficient in MSA (all women[5]; mean age: 19; *SD* = 1.41) were recruited from the United Arab Emirates University (UAEU). Participants were offered both course credit and 40 AED (~15 USD) in cash compensation for their participation. Prior to beginning the experiment, participants completed a detailed language history questionnaire in which they rated their proficiency in listening, speaking, reading and writing (scale from 1 to 7) in MSA and all their other known languages (see Appendix D for the full list of questions). These questions were adapted from the Language History Questionnaire 3.0 (Li et al., 2020), which is used to evaluate language proficiency for speakers of multiple languages. This questionnaire differed from our previous questionnaires in that participants were explicitly asked to rate their proficiency in MSA across various spectrums, rather than simply identifying themselves as a native Arabic speaker. Participants were considered proficient in MSA if they (1) selected Arabic as their native language, and (2) scored their proficiency in each area for MSA at 4 or higher. One participant was excluded for selecting English as their native language. Further, we established an a priori criterion to exclude any participant who scored lower than 75% accuracy on comprehension questions on filler items, but all participants performed above this criterion. One final participant was excluded due to a technical error during their experiment.

4.2.2. Materials

We used the 40 original items from our self-paced reading study, then created an additional 45 items following the same design (see Section 3.2.2). A norming study was conducted on our 45 new items to

[5] The UAEU campus is segregated by gender, so we were limited to recruiting and testing only female participants.

confirm that the subject and object of each relative clause were equally plausible in both clause conditions. Native Arabic speakers (N = 80; 24 women, 54 men, 2 not reported; mean age: 32; SD = 10.64) were recruited through Prolific and asked to rate the plausibility of each sentence on a Likert scale (1 = highly implausible, 7 = highly plausible). Plausibility ratings were collected for both the full stimuli sentences and the relative clauses as simplified transitive sentences. Five stimuli were excluded after a paired t-test revealed substantial discrepancies between plausibility ratings in the SRC and ORC conditions for those items. The mean plausibility rating for full stimuli sentences for the remaining items was 5.77 (SD = 1.63) for SRCs and 5.76 (SD = 1.64) for ORCs, and the mean plausibility rating for the simplified transitive sentences was 6.38 (SD = 1.27) for SRCs and 6.39 (SD = 1.23) for ORCs. The 40 remaining new stimuli combined with the 40 original items resulted in 80 total stimuli. We also included 80 unrelated filler sentences for a total of 160 sentences. Experimental items were split into four lists and counterbalanced within and across lists in a Latin square design. Participants were randomly assigned to one of four lists, and the order of sentences within each list was randomized separately for each participant.

Comprehension questions targeting comprehension of the relative clause appeared after all stimuli and general comprehension questions appeared after all filler items. Our previous comprehension questions had simple "Yes" or "No" options; however, the findings from our comprehension question analysis suggested that this framing of the comprehension questions may have biased the results. To mitigate this possible bias, we offered full-sentence options in the same forced-choice structure (e.g., "Which of the following happened?" (a) the reporter attacked the senator, (b) the senator attacked the reporter). We also randomized whether the correct or incorrect answer was presented first and counterbalanced by item type.

4.2.3. Apparatus and Procedure

The experiment took place at the UAEU in the Department of Cognitive Science's eye tracking lab using the Eyelink 1000 Plus eye tracker (SR Research). Right eye gaze movements were recorded via a high-

speed 35 mm lens on a desktop mount at a sampling rate of 1000 Hz. Participants' head movements were stabilized using a head stabilization tower. Sentences were displayed on a high definition (1920 x 1080 pixels) 24" BENQ ZOWIE XL 2430 monitor at 80 cm viewing distance. Text was presented in 20-point Times New Roman, a proportionally spaced typeface.

Participants were tested individually in a quiet, isolated room. Instructions were given verbally in English and were presented in both English and SA text on the screen. The language of instruction at UAEU is English, so a baseline proficiency in English was assumed. The eye tracker was calibrated and validated using the default 9-point grid calibration. Participants completed two practice trials and were allowed to ask questions before proceeding with the experiment.

The experiment was divided into eight blocks of 20 items each to allow for breaks throughout the experiment. Calibration accuracy was assessed before the start of each block and the eye tracker was re-calibrated as necessary. At the beginning of each trial, a right-aligned asterisk was placed at the onset site of the first letter of the sentence. Once the participant fixated for at least 800 ms, the sentence appeared and replaced the asterisk. Participants pressed the spacebar on the keyboard once they were finished reading the sentence, then used the mouse to click and select the correct comprehension question answer on the following screen. Completion of the experiment took about 60 minutes for each participant.

4.3. Analysis and Results

Data were cleaned using SR Research's Data Viewer. Following standard procedures, fixations that were less than 80 ms and within one character of each other were merged, and remaining fixations less than 80 ms or longer than 1,200 ms were excluded. We also excluded trials where significant track loss occurred and fixations where a blink occurred on the target word. This resulted in a total data loss of 5.8%.

After cleaning our data and before conducting our analyses, we investigated the comprehension question accuracy rates for SRCs versus ORCs. We found an overall accuracy rate of 84.4%, with a 93.5% accuracy for SRCs compared to only 75.4% accuracy for ORCs. In terms of number of trials, we

observed 1,747 correct SRC trials, 1,413 correct ORC trials, and 462 incorrect ORC trials. We also calculated accuracy by participant and found substantial variance by individual. ORC scores ranged from 10% to 100%; seven participants (out of 47) scored less than 50%, while only two participants scored 100%. Conversely, the lowest score on SRCs was 70%, with the second lowest score being 75%, and ten participants scored 100%. A distribution of the average scores by participant and clause type are shown in Figure 4.1.

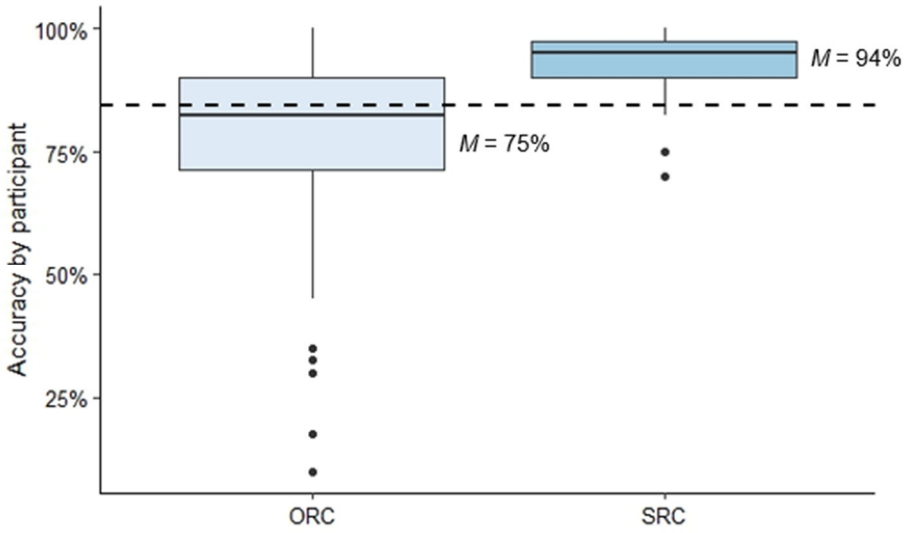

Figure 4.1: Box plot of average accuracy by participant for each clause type. The horizontal dashed line represents the overall mean of 84%.

For our analysis, interest areas were divided into regions as previously illustrated in Figure 2.2. Up to 3 spillover regions (one word each) were analyzed when sentences were long enough. All sentences had at least one spillover region. For each region, we calculated the following eye tracking metrics using the Get Reading Measures package from SR Research: *first fixation duration* (the duration of the first fixation on a region), *first pass duration* (the sum of all first pass fixations before leaving a region for the first time), *go-past time* (the sum of all first pass fixations on a region, including any time spent reading previous material, until progressively leaving the region for the first time), *total fixation duration* (the sum of all fixations on a region), *first pass regression* (a binary measure indicating whether the reader's

first pass through a region ended with a regressive saccade to an earlier part of the sentence), and *first pass skip* (a binary measure indicating whether a reader skipped a region on first-pass reading).

To evaluate the statistical significance of our data, linear mixed-effects regression models were fit for numeric reading measures and logistic mixed-effects regression models were fit for binary measures. We fit individual models for each eye tracking metric in each interest area for each analysis (Table 4.1). The models included sum-coded fixed effects for Clause Type (ORC: 1, SRC: -1 for Cor ORC vs. Cor SRC and Incor ORC vs. Cor SRC models) or Correctness (Correct: 1, Incorrect: -1 for Incor ORC vs. Cor ORC models), where Correctness was a measure of participant-specific accuracy by trial. We also included control predictors of Word Length (numeric) and Trial Index (numeric). We used the maximal random effects structure justified by the design, resulting in random intercepts for Participant and Item, and random slopes by Clause Type for both Participant and Item. Model estimates for main effects are reported in Table 4.2.

4.3.1. Correctly interpreted ORCs versus correctly interpreted SRCs

We found no significant effects for Clause Type at the relative clause verb for any fixation metric. However, we did find significant effects at the relative clause NP for go-past time, total fixation duration, and first pass regressions. This indicates that ORCs were associated with significantly longer reading times and higher regression rates. We also found a significant effect for Clause Type for total fixation duration at the matrix verb and spillover region 1, such that ORCs were read faster in these regions.

4.3.2. Misinterpreted ORCs versus correctly interpreted ORCs

We found no significant effects for Correctness at the relative clause verb for any fixation metric, but again found significant effects at the relative clause NP for go-past time, total fixation duration, and first pass regressions. This demonstrates that readers spent more time reading and made more regressions

Table 4.2: Linear and logistic mixed-effects model estimates of the dependent variable, Clause Type or Correctness, on all fixation metrics by region, including SE estimates and CrIs. The matrix NP region does not have an estimate for p(regress) as it is the first region in each sentence and a regressive saccade would be impossible. The % > 0 column shows the percent of the sampled posterior distribution that is over 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates that are bolded are significant.

Fixation Metrics by Region	**Cor ORC vs. Cor SRC**				**Incor ORC vs. Cor ORC**				**Incor ORC vs. Cor SRC**			
	β	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**
matrix NP												
first fixation	0.90	2.70	[-4.43, 6.23]	63	0.81	4.02	[-6.96, 8.81]	58	1.43	4.23	[-7.06, 9.67]	64
first pass	4.29	5.78	[-7.08, 15.75]	77	-2.68	9.98	[-22.03, 16.94]	39	1.77	8.70	[-15.29, 18.87]	58
go-past	4.31	5.80	[-7.08, 15.59]	77	-2.70	9.94	[-22.23, 16.92]	39	1.78	8.81	[-15.41, 19.05]	58
total duration	26.61	21.69	[-16.23, 69.18]	89	-21.34	28.68	[-77.08, 35.06]	23	-4.56	21.81	[-46.98, 38.51]	42
p(skip)	-0.06	0.07	[-0.20, 0.09]	22	0.00	0.10	[-0.20, 0.19]	51	-0.08	0.11	[-0.32, 0.14]	23
RC pronoun												
first fixation	-0.27	2.33	[-4.86, 4.30]	45	-0.26	4.05	[-8.15, 7.59]	48	0.95	3.56	[-5.99, 7.97]	61
first pass	2.37	3.09	[-3.70, 8.50]	78	-2.42	5.36	[-12.83, 8.12]	32	1.90	4.36	[-6.64, 10.41]	67
go-past	2.52	5.48	[-8.17, 13.21]	68	-10.69	9.67	[-29.52, 8.30]	13	-10.12	7.38	[-24.53, 4.42]	8
total duration	5.88	14.01	[-21.47, 33.65]	67	-4.99	19.18	[-42.40, 33.12]	39	3.24	12.95	[-22.27, 28.72]	59
p(regress)	-0.13	0.08	[-0.29, 0.03]	6	-0.11	0.14	[-0.40, 0.16]	21	**-0.23**	**0.13**	**[-0.49, 0.01]**	**3**
p(skip)	-0.02	0.04	[-0.10, 0.07]	36	0.05	0.08	[-0.10, 0.20]	74	0.02	0.07	[-0.12, 0.15]	60
RC verb												
first fixation	3.47	3.35	[-3.20, 9.94]	85	2.33	5.31	[-8.09, 12.81]	67	-1.01	5.57	[-11.76, 10.16]	42
first pass	3.56	7.20	[-10.60, 17.69]	69	-7.01	9.73	[-25.85, 12.38]	23	-6.73	8.88	[-24.01, 11.03]	22
go-past	5.00	12.95	[-20.23, 30.30]	65	-2.68	16.00	[-34.09, 28.51]	43	17.04	20.11	[-21.51, 57.79]	80
total duration	33.52	26.40	[-19.24, 84.15]	90	-3.21	32.15	[-65.70, 61.83]	45	44.06	39.45	[-32.30, 122.93]	87
p(regress)	0.02	0.08	[-0.13, 0.18]	61	-0.13	0.14	[-0.43, 0.13]	18	0.02	0.14	[-0.28, 0.28]	59
p(skip)	0.00	0.07	[-0.13, 0.13]	51	-0.06	0.11	[-0.29, 0.14]	28	0.06	0.10	[-0.15, 0.25]	73
RC NP												
first fixation	3.54	3.19	[-2.77, 9.76]	87	**-12.67**	**5.88**	**[-24.40, -1.28]**	**1**	-8.12	5.22	[-18.43, 2.10]	6
first pass	4.44	7.67	[-10.67, 19.60]	72	-5.60	12.64	[-31.05, 18.92]	33	-5.71	9.78	[-25.00, 13.40]	27
go-past	**38.66**	**20.82**	**[-2.65, 79.61]**	**97**	**-59.94**	**26.28**	**[-110.93, -8.58]**	**1**	-24.33	22.20	[-68.44, 19.02]	13
total duration	**59.97**	**17.83**	**[24.67, 94.84]**	**100**	**-76.91**	**25.77**	**[-127.94, -26.58]**	**0**	-27.90	23.12	[-73.32, 17.96]	11
p(regress)	**0.11**	**0.06**	**[0.00, 0.23]**	**98**	**-0.26**	**0.10**	**[-0.46, -0.07]**	**0**	-0.13	0.10	[-0.35, 0.06]	10
p(skip)	0.01	0.08	[-0.15, 0.16]	57	-0.08	0.16	[-0.42, 0.21]	31	-0.05	0.17	[-0.41, 0.25]	40
matrix verb												
first fixation	-4.28	2.87	[-9.90, 1.38]	7	4.29	4.97	[-5.49, 13.95]	81	0.79	4.47	[-8.06, 9.42]	57
first pass	-5.48	3.73	[-12.77, 1.87]	7	7.46	6.63	[-5.50, 20.51]	87	-0.77	6.12	[-13.02, 11.08]	45
go-past	0.93	13.14	[-24.91, 26.81]	52	42.62	40.42	[-33.28, 126.10]	86	42.85	47.02	[-49.49, 137.91]	82
total duration	**-15.67**	**7.51**	**[-30.45, -0.87]**	**2**	15.85	13.86	[-11.16, 43.47]	88	-10.96	12.39	[-35.10, 13.68]	19
p(regress)	-0.05	0.07	[-0.19, 0.08]	21	-0.05	0.12	[-0.30, 0.18]	33	-0.18	0.13	[-0.46, 0.08]	9
p(skip)	0.06	0.05	[-0.03, 0.16]	90	-0.11	0.09	[-0.30, 0.06]	11	-0.06	0.08	[-0.23, 0.09]	21
spillover 1												
first fixation	-1.57	2.57	[-6.59, 3.49]	27	-1.90	4.77	[-11.29, 7.38]	35	-1.76	4.53	[-10.84, 7.16]	35
first pass	-4.50	5.74	[-15.61, 6.90]	21	1.93	9.61	[-16.99, 21.19]	58	-0.85	6.65	[-13.75, 12.36]	44
go-past	9.05	19.23	[-28.36, 46.29]	68	-40.76	34.86	[-109.22, 27.44]	12	-20.53	28.06	[-76.38, 34.01]	23
total duration	**-13.60**	**6.99**	**[-27.30, 0.24]**	**3**	13.22	16.11	[-19.07, 44.70]	80	2.86	12.38	[-21.20, 27.32]	59
p(regress)	-0.03	0.07	[-0.16, 0.10]	30	-0.01	0.12	[-0.26, 0.23]	50	-0.09	0.12	[-0.33, 0.13]	21
p(skip)	0.06	0.05	[-0.03, 0.16]	91	-0.08	0.09	[-0.25, 0.09]	18	-0.03	0.08	[-0.18, 0.12]	36

spillover 2												
first fixation	1.96	2.75	[-3.50, 7.36]	77	4.23	5.10	[-5.80, 14.30]	80	**8.84**	**4.10**	**[0.73, 16.84]**	**98**
first pass	-0.11	3.90	[-7.76, 7.59]	49	2.59	6.68	[-10.33, 15.87]	65	3.60	6.02	[-8.26, 15.37]	72
go-past	23.54	26.27	[-27.76, 75.12]	81	-2.33	81.09	[-161.44, 156.20]	49	23.07	83.45	[-140.78, 186.04]	61
total duration	1.92	6.50	[-10.81, 14.73]	61	6.12	11.85	[-17.24, 29.29]	70	11.54	10.25	[-8.61, 31.68]	87
p(regress)	0.01	0.07	[-0.13, 0.14]	54	**-0.28**	**0.14**	**[-0.57, -0.01]**	**2**	**-0.21**	**0.12**	**[-0.45, 0.02]**	**4**
p(skip)	0.00	0.05	[-0.10, 0.10]	53	-0.08	0.11	[-0.30, 0.12]	22	-0.12	0.10	[-0.31, 0.06]	10
spillover 3												
first fixation	-2.45	2.94	[-8.13, 3.45]	20	6.73	5.44	[-4.04, 17.46]	89	5.93	4.74	[-3.42, 15.24]	90
first pass	1.79	4.48	[-6.98, 10.66]	65	5.68	8.95	[-11.85, 23.34]	74	10.52	7.61	[-4.19, 25.87]	92
go-past	33.65	43.64	[-52.70, 118.51]	78	-52.93	70.59	[-191.44, 84.75]	23	-24.85	50.69	[-123.56, 75.63]	31
total duration	4.33	8.10	[-11.82, 20.26]	71	7.25	15.62	[-23.00, 38.48]	68	14.47	11.37	[-7.68, 37.01]	90
p(regress)	0.04	0.10	[-0.16, 0.24]	67	**-0.31**	**0.17**	**[-0.64, 0.01]**	**3**	-0.26	0.17	[-0.62, 0.07]	6
p(skip)	0.06	0.06	[-0.05, 0.17]	88	-0.12	0.11	[-0.34, 0.10]	15	-0.03	0.11	[-0.24, 0.17]	41

away from the relative clause NP when they correctly interpreted ORCs versus when they misinterpreted them. There was also a significant effect for Correctness for first pass regressions at spillover regions 2 and 3. This effect suggests that participants regress to re-read and confirm their interpretation of the relative clause when arriving at the correct interpretation of the sentence (Christianson et al., 2017).

While comparing misinterpreted versus correct ORCs, we also wanted to look at fixation metrics on the resumptive pronoun clitic specifically. The analyses at the relative clause verb included fixations on the resumptive pronoun, and in general we do not expect the resumptive pronoun to be fixated directly in every trial. Nonetheless, we aimed to investigate whether fixation metrics on the resumptive pronoun clitic directly had any effect on the correct comprehension of ORCs. Average fixation metrics by clitic (masculine singular (M.Sg)– one character or feminine singular (F.Sg) – two characters) and correctness (incorrect or correct ORC) are reported in Table 4.3. Numerically, both M.Sg and F.Sg clitics were fixated for less time in incorrect ORCs than in correct ORCs, and clitics were more likely to be fully skipped for incorrect ORCs than correct ORCs. On the other hand, the probability of a regression was consistent across clitics and correctness with the exception of F.Sg for incorrect ORCs, which was much less likely to trigger a regression. A supplementary analysis investigating general fixation patterns on clitics in the stimuli is included in Appendix E.

Table 4.3: Average fixation metrics by clitic and correctness. The masculine singular resumptive object pronoun clitic in Arabic is one character in length while the feminine singular clitic is two characters in length.

	Incorrect ORCs		Correct ORCs	
Fixation Metrics	**M.Sg (ـه)**	**F.Sg (ـها)**	**M.Sg (ـه)**	**F.Sg (ـها)**
first fixation	53.85	58.89	45.52	64.88
first pass	53.85	63.10	46.20	69.58
go-past	72.92	89.32	62.60	110.28
total duration	142.52	153.70	175.64	234.84
p(regress)	0.27	0.16	0.25	0.26
p(first pass skip)	0.80	0.79	0.83	0.76
p(overall skip)	0.62	0.62	0.55	0.52

To test the statistical significance of these differences, we ran additional regression models in this region following the same design outlined above. The model estimates for main effects are reported in Table 4.4. While Correctness did not have a significant effect on any fixation metric in this region, Clitic Length had a significant effect on first pass skip rates (β = -0.36; *SE* = 0.13; CrI = [-0.62, -0.09]), such that two-character clitics (F.Sg) were significantly less likely to be skipped than one-character clitics (M.Sg).

Table 4.4: Linear and logistic mixed-effects model estimates of the dependent variable, Correctness, on all fixation metrics in the resumptive pronoun clitic region, including SE estimates and CrIs. Estimates marked with an asterisk are significant.

	Incor ORC vs. Cor ORC		
Fixation Metrics	**β**	***SE***	**CrI**
RP clitic			
first fixation	3.84	4.08	[-4.19, 11.85]
first pass	3.13	4.37	[-5.53, 11.81]
go-past	1.43	7.06	[-12.49, 15.35]
total duration	-7.92	10.35	[-28.24, 12.18]
p(regress)	-0.36	0.38	[-1.28, 0.21]
p(skip)	-0.02	0.09	[-0.20, 0.17]

4.3.3. Misinterpreted ORCs versus correctly interpreted SRCs

We found no significant effects for Clause Type at the relative clause verb or the relative clause NP for any fixation metric. We did, however, find significant effects for Clause Type for first pass regressions at the relative pronoun, such that readers were less likely to make a first pass regression at the relative

pronoun for misinterpreted ORCs than for correct SRCs. This finding is surprising because, at this point in the sentence, the SRC and ORC versions of each stimulus are identical and the reader has not yet encountered any disambiguating information about the clause type. Notably, this same effect was near but not quite significant in the model that compared correct ORCs to correct SRCs, and the effect for Correctness for misinterpreted versus correct ORCs in the same region was also insignificant.

Finally, we found a significant effect for Clause Type in spillover region 2 for first fixation duration and first pass regressions. Misinterpreted ORCs were read longer on first pass than correct SRCs in this region, but had fewer first pass regressions. This effect for first pass regressions matches what we found in this region when comparing misinterpreted and correct ORCs, and further suggests that regressions may be made to re-read and confirm veridical interpretations.

4.3.4. Re-investigating the relative clause verb (again)

Similar to our self-paced reading experiment, despite finding no significant effects on the relative clause verb for any of our analyses, we wanted to determine whether readers were actually registering the resumptive pronoun clitic on the ORC verb and paying any sort of additional processing cost for reading it. All of our models included Word Length as a control variable, so it was possible that this extra processing cost was present but proportional to length. To investigate this possibility, we re-ran all our models in the relative clause verb region but excluded Word Length as a predictor variable. Model estimates are reported in Table 4.5.

Model estimates showed significant effects by Clause Type at the relative clause verb when comparing correct ORCs to correct SRCs. Correct ORCs had longer first fixation, first pass, go-past, and total duration times than correct SRCs. This same pattern held for Correctness for misinterpreted ORCs compared to correct SRCs: misinterpreted ORCs also had longer first fixation, first pass, go-past and total duration times than correct SRCs. On the other hand, there was no significant difference in reading times between misinterpreted and correct ORCs. These data suggest that readers do register the resumptive

Table 4.5: Linear and logistic mixed-effects model estimates of the dependent variable, Clause Type or Correctness, without the Word Length control predictor on all fixation metrics in the relative clause verb region, including SE estimates and CrIs. The % > 0 column shows the percent of the sampled posterior distribution that is over 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates marked with an asterisk are significant.

Fixation	**Cor ORC vs. Cor SRC**				**Incor ORC vs. Cor ORC**				**Incor ORC vs. Cor SRC**			
Metrics	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**
RC verb												
first fixation	11.92*	2.95	[6.18, 17.66]	100	2.75	5.32	[-7.67, 13.12]	30	9.08*	5.16	[-1.11, 19.25]	96
first pass	42.33*	6.68	[29.03, 55.47]	100	-7.50	9.88	[-27.07, 11.66]	78	32.86*	8.05	[17.05, 48.74]	100
go-past	54.81*	11.49	[32.51, 77.76]	100	-2.64	15.87	[-33.37, 29.03]	57	57.85*	18.81	[21.67, 95.52]	100
total duration	141.15*	23.62	[94.92, 187.32]	100	-7.55	32.37	[-70.32, 57.77]	60	142.84*	35.93	[72.72, 214.65]	100
p(regress)	0.04	0.07	[-0.10, 0.18]	71	-0.12	0.14	[-0.41, 0.13]	82	-0.07	0.14	[-0.36, 0.18]	33
p(skip)	-0.20*	0.06	[-0.32, -0.09]	0	-0.07	0.11	[-0.30, 0.13]	74	-0.13	0.09	[-0.32, 0.05]	8

pronoun clitic on the ORC verb and read ORC verbs longer than SRC verbs, regardless of whether they correctly interpret the ORC. However, this increased reading time at the ORC verb is proportional to its increased length due to the inclusion of the resumptive pronoun clitic.

4.3.5. Participant accuracy over time

While collecting data for the eye tracking experiment, one author received a few comments from different participants that they learned that there were certain parts of the sentence that they needed to pay attention to in order to answer the comprehension questions correctly. These comments led us to question whether participants were perhaps tailoring their reading and comprehension strategy to the experiment, and improving comprehension question accuracy over the course of the experiment. We decided to conduct an exploratory analysis to see if there was a general trend of higher comprehension question accuracy in the second half of the experiment than in the first, and whether participants used distinct reading strategies in either half.

The number of correct versus incorrect items by clause type for the first and second halves of the experiment are reported in Table 4.6. Accuracy rates for SRCs were very similar in the first and second halves of the experiment; however, there appeared to be a substantial improvement for ORCs in the second half of the experiment. Average ORC accuracy in the first half was only 68.1% compared to an

average 82.4% accuracy in the second half. We also considered whether there were substantial differences by participant in comprehension question accuracy improvement. A plot showing average ORC accuracy by participant for the first and second halves of the experiment is included in Appendix F.

Table 4.6: Number of correct and incorrect trials, plus their respective percentages, by clause type for the first and second halves of the experiment.

	First Half		Second Half	
Clause Type	**Correct**	**Incorrect**	**Correct**	**Incorrect**
SRC	884 (93.7%)	59 (6.3%)	863 (93.2%)	63 (6.8%)
ORC	628 (68.1%)	294 (31.9%)	785 (82.4%)	168 (17.6%)

To investigate whether there were significant differences in eye movement behaviors in the first and second halves, we ran additional regression models for all three of our analyses following the same design outlined above, but fit separate models for the data collected during the first half of the experiment versus the second half (e.g., first fixation duration for Cor ORC vs. Cor SRC in the relative clause verb region had a first-half model and a second-half model). Selected model estimates for the Cor ORC vs. Cor SRC analysis in the relative clause verb and NP regions are reported in Table 4.7.

Table 4.7: Linear and logistic mixed-effects model estimates for Cor ORC vs. Cor SRC of the dependent variable, Clause Type, on all fixation metrics in the relative clause verb and NP region for trials in the first versus second half of the experiment, including SE estimates and CrIs. The % > 0 column shows the percent of the sampled posterior distribution that is over 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates marked with an asterisk are significant.

Fixation Metrics	**First Half**				**Second Half**			
by Region	**β**	***SE***	**CrI**	**% > 0**	**β**	***SE***	**CrI**	**% > 0**
RC verb								
first fixation	-0.26	4.58	[-9.26, 8.74]	48	9.00*	4.83	[-0.48, 18.46]	97
first pass	4.28	9.71	[-14.72, 23.55]	67	5.61	9.53	[-13.12, 24.18]	73
go-past	-1.03	16.30	[-33.20, 30.68]	48	15.69	15.67	[-15.08, 46.75]	84
total duration	40.91	36.43	[-30.92, 112.86]	87	21.39	26.82	[-31.50, 73.30]	79
p(regress)	0.00	0.10	[-0.20, 0.20]	48	0.08	0.12	[-0.17, 0.32]	73
p(skip)	0.00	0.10	[-0.21, 0.19]	51	-0.06	0.09	[-0.24, 0.12]	25
RC NP								
first fixation	6.21	4.61	[-2.95, 15.18]	91	0.96	4.67	[-8.33, 10.14]	58
first pass	0.42	8.23	[-15.57, 16.79]	52	6.41	10.62	[-14.62, 27.25]	73
go-past	44.89*	23.18	[-0.96, 89.77]	97	31.88	24.45	[-15.57, 80.38]	90
total duration	61.84*	29.04	[4.90, 118.49]	98	42.67*	21.42	[-0.39, 84.93]	97
p(regress)	0.16*	0.07	[0.02, 0.30]	99	0.07	0.08	[-0.10, 0.24]	80
p(skip)	-0.06	0.11	[-0.28, 0.17]	31	0.07	0.12	[-0.17, 0.29]	73

For the first half of items, model estimates showed significant effects by Clause Type at the relative clause NP, such that go-past and total duration times were longer for correct ORCs than correct SRCs. There were no significant differences by Clause Type at the relative clause verb. These estimates all matched our findings from our main analyses in these regions. On the other hand, for the second half of items, model estimates showed a significant effect by Clause Type for first fixation duration at the relative clause verb, such that correct ORCs had longer first fixations than correct SRCs in that region. There was also a significant effect for Clause Type on total duration at the relative clause NP, similar to what was seen in the first half.

These findings suggest that participants may have adjusted their reading strategies from the first to the second half of the experiment. Whereas there was no significant difference in early or late processing measures at the relative clause verb region in the first half, participants appeared to pay closer to the relative clause verb upon first fixation in the second half, perhaps to identify the resumptive object pronoun clitic that was indicative of an ORC interpretation. Participants then paid an integration cost at the relative clause NP regardless of their reading strategy at the verb. It is important to note that our experiment did not have any specific manipulation that would encourage readers to change their reading strategy half-way through the experiment, so these findings are merely exploratory; however, the results bring up interesting considerations that may be of use for designing future experiments of this type.

4.4. Discussion

This study set out to investigate the cause of misinterpretations while processing Arabic SRCs and ORCs through an eye tracking study. We first tested whether SRCs or ORCs were harder to process in Arabic in an effort to corroborate previous findings. To answer this question, we analyzed differences during veridical processing, between SRC and ORC trials with correct comprehension question answers. Based on the results of our corpus analysis, both memory- and expectation-based theories predict the relative clause verb as the locus of distinguishing processing behavior: memory-based theories predict comparable processing difficulty due to the presence of an object resumptive pronoun clitic, while expectation-based

theories predict increased processing difficulty for ORCs, the less frequent structure, upon probabilistically disambiguating the clause at the relative clause verb. We further tested previous findings from English relative clauses that these processing difficulties manifest in distinct behaviors (Staub, 2010), with longer go-past times indicating difficulty from memory constraints and increased regressive saccades indicating difficulty from violated expectations.

Our results showed that ORCs were read significantly more slowly than SRCs overall, in line with our previous findings and in support of expectation-based theories. We specifically found that ORCs had significantly longer go-past times, total fixation durations, and increased regressive saccades. According to Staub (2010), this would indicate processing difficulty from both memory limitations and violated expectations. Our findings contradict these predictions – our stimulus design indicates that there should only be expectation-based difficulty, but the observed difficulty manifested in behaviors attributed by Staub to both expectation- and memory-based difficulty.

The results of our corpus analysis indicated that the relative clause verb should be the probabilistic disambiguating region and the predicted site of processing difficulty. However, our results showed no significant difference in processing by clause type at the relative clause verb, following the results from our self-paced reading task. Readers do spend more time reading the relative clause verb for ORCs than SRCs, but this is proportional to the added length from the resumptive object pronoun clitic. Rather, we found significant differences in processing at the relative clause NP, also in line with previous findings from our self-paced reading task.

Taken together, these results suggest that readers pay a processing cost when integrating the relative clause NP in the globally less-expected ORC structure, even though they had previously received probabilistic disambiguating information. These results are not predicted by strict, incremental expectation-based processing theories. Expectation-based theories operationalized through surprisal predict that there should be more processing difficulty for ORCs, but that this cost would have been paid at the relative clause verb where the reader can probabilistically disambiguate between an SRC and an

ORC. Theories that consider effects from both memory constraints and expectations, such as noisy- or lossy-context surprisal (Futrell et al., 2020; Futrell & Levy, 2017) also do not predict these effects, as both memory- and expectation-based theories suggest that waiting until the relative clause NP should be less advantageous and incur more processing difficulty for the comprehender.

One possible explanation for these results lies in the temporary ambiguity of relative clauses in MSA (see Section 1.4.2). We received a few comments from native and non-native speakers that readers could initially be interpreting the relative clause verb with the resumptive object pronoun as a SRC with a resumptive object pronoun and null object noun phrase. In this case, the reader would maintain uncertainty about the interpretation of the clause until they encountered the relative clause NP, despite seeing a resumptive pronoun clitic, and then pay the processing cost of updating their expectations in that region. We derived our predictions from expectation-based theories from our corpus analyses, which showed that while this structure does occur in SRCs in MSA, it is much less common than both the standard SRC construction *and* standard ORC construction. If readers are maintaining an SRC interpretation upon reading the relative clause verb with a resumptive pronoun, they would be relying on a global expectation for SRCs over ORCs, despite the fact that more granular expectations should favor ORCs when the relative clause verb has a resumptive pronoun. We cannot definitively say whether this is the case or whether readers do probabilistically switch to an ORC interpretation at the relative clause verb, yet still pay a processing cost at the relative clause NP. But in either case, our findings show that readers are not doing the most granular, incremental processing possible while reading Arabic relative clauses.

We also asked whether misinterpretations of ORCs were caused by readers misreading ORCs as SRCs by missing the resumptive pronoun clitic, or correctly reading ORCs and instead accepting a noisy but preferred SRC interpretation. To answer this question, we analyzed the differences in processing behaviors in misinterpreted ORC trials compared to both correct ORC and correct SRC trials. In the case of misreading, we expected to see no significant difference between incorrect ORCs and correct SRCs at

the relative clause verb, but a significant difference between incorrect and correct ORCs. This would indicate that incorrect ORCs were read as SRCs by missing the resumptive pronoun clitic, and distinctly from incorrect ORCs. In the case of good-enough or noisy-channel processing, we expected to see the opposite: a significant difference between incorrect ORCs and correct SRCs at the relative clause verb, and no significant difference between incorrect and correct ORCs. This would indicate that readers register the resumptive pronoun clitic on the verb for incorrect ORCs, yet later accept a noisy, incorrect interpretation.

We found no significant differences in reading times or regression rates between misinterpreted and correctly interpreted ORC trials at the relative clause verb. This suggests that readers were not misreading the verb when misinterpreting ORCs, as incorrect ORCs behave like correct ORCs. Rather, differences between incorrect and correct ORCs manifested at the relative clause NP, where correct ORCs had longer reading times and higher regression rates. Notably, this is the same region where correct ORCs incur the most processing difficulty relative to correct SRCs. On the other hand, incorrect ORCs and correct SRCs had no significant differences at the relative clause verb based on cause type; however, incorrect ORCs were read longer than SRCs proportional to the added length from the resumptive pronoun clitic. Further, incorrect ORCs and correct SRCs had no significant difference at the relative clause NP. So, incorrect ORCs behave similarly to correct ORCs at the relative clause verb, but similarly to correct SRCs at the relative clause NP.

Overall, our results show that the locus of processing difficulty is at the integration of the relative clause NP, where readers can definitively know whether they are reading an SRC or an ORC, rather than at the relative clause verb, where readers can probabilistically disambiguate between the two. Accepting a noisy SRC interpretation of an ORC skips the integration cost at the relative clause NP, resulting in comparable processing difficulty between incorrect ORCs and correct SRCs in this region. These results thus support noisy-channel processing over misreading as an explanation for why some ORCs are misinterpreted as SRCs.

Chapter 5

Experiment 4: Eye tracking part 2

5.1. Introduction

Following the findings from Experiment 3, the goal of our final experiment was to test how grammatical and/or orthographic cues affect a reader's willingness to accept a good-enough or noisy interpretation in Arabic ORCs. Namely, we ask the question, does strengthening grammatical cues for a low-frequency interpretation reduce expectations of a high-frequency interpretation in a good-enough/noisy-channel processing framework?

Prior existing literature on agreement marking suggests that these grammatical cues are indeed important, especially in languages with flexible word order: while these markers may be superfluous and redundant in strict SVO word-order languages, it is beneficial in verb-initial and verb-final languages as a lack of agreement marking results in delays in argument processing (Hawkins, 2004; Sinnemäki, 2010). Given that the relative clauses in question here are all verb-initial, agreement marking should facilitate processing. Further, the resumptive pronoun clitics in Arabic ORCs are grammaticalized, which has been shown to improve processing times (Meltzer-Asscher, 2021). However, some prior work on agreement marking and good-enough/noisy-channel processing suggests that these additional cues may not make a significant impact on a reader's likelihood of accepting a good-enough or noisy interpretation. For

example, Frazier (1987) compared preferences for SRCs versus ORCs in Dutch and found that readers had a lingering bias for SRCs even when grammatical agreement provided evidence in favor of an ORC interpretation. Thus, readers were willing to accept a grammatical mismatch with a higher-frequency structure over a grammatical match with a lower-frequency structure. Notably, Dutch has strict SVO word order, unlike Arabic which has flexible word order. On the other hand, previous research on this topic in Hebrew found mixed results. Keshev & Meltzer-Asscher (2021) tested whether creating grammatical mismatches between NPs in SRCs versus ORCs affected the reader's willingness to accept a noisy interpretation and tested this effect in two types of ORC structures – one that was rare, and one that was more common. When comparing a rare ORC structure to a frequent SRC structure, readers ignored grammatical cues in favor of an ORC interpretation; however, when comparing a more common ORC structure to a frequent SRC structure, readers were *not* as willing to ignore these cues, and were less likely to accept the noisy SRC interpretation. Since Arabic is typologically similar to Hebrew, we wanted to investigate whether we would find similar results in our study.

We tested this question by conducting a similar eye-tracking task as outlined in Experiment 3, while manipulating the grammatical number and gender of the matrix and relative clause nouns. In Experiment 3, the relative clause verb and resumptive object pronoun agreed grammatically with both the matrix and relative clause verb, meaning that the resumptive object pronoun clitic provides the unique signal of an ORC interpretation. By manipulating gender and number to create a mismatch between the matrix and relative clause nouns, the resumptive pronoun clitic agrees only with the matrix clause subject and the relative clause verb agrees only with the relative clause noun phrase, creating additional signals for an ORC interpretation (Keshev & Meltzer-Asscher, 2021). Creating these additional grammatical manipulations also allowed us to control for the temporary ambiguity of an ORC at the relative clause verb: when the matrix and relative clause NPs are matched for number and gender, an ORC with an RP can be temporarily interpreted as an SRC with a RP and null object. When these NPs no longer match in

number or gender, this ambiguity at the relative clause verb goes away as the verb *only* agrees grammatically with the relative clause NP, and not the matrix clause NP.

We revised our stimuli to have three different grammatical conditions: a Match condition, a Single Mismatch condition, and a Double Mismatch condition (Figure 5.1). The Match condition mirrored the stimuli from Experiment 3, where the matrix clause NP and relative clause NP matched in both grammatical number (singular) and gender (masculine or feminine). In the Match condition, the ambiguity of the clause is maintained until the reader encounters the relative clause NP, as the conjugation of the relative clause verb agrees with both the matrix and relative clause noun. In the Single Mismatch condition, we changed the number of the matrix clause NP while the relative clause NP remained unchanged. For example, if the relative clause NP was masculine singular, the matrix clause NP was masculine plural. In this case, in the ORC condition, the relative clause verb now only grammatically agrees with the relative clause NP, and the resumptive pronoun only grammatically agrees with the matrix clause NP. Finally, in the Double Mismatch condition, we changed both the gender and the number of the matrix clause NP while the relative clause again remained unchanged. For example, if the relative clause NP was masculine singular, the matrix clause NP was feminine plural. In this case, the relative clause verb and resumptive pronoun in the ORC condition would again only agree with their referent NPs, but the double mismatch provides an additional grammatical cue in favor of a veridical interpretation.

Match:	The bus driver (**M.sg**) who the kid (**M.sg**) followed (him.**M.sg**) wondered about the location of the hotel.	سائق الحافلة الذي تبعه الطفل تساءل عن موقع الفندق.
Single mismatch: (number)	The bus drivers (**M.pl**) who the kid (**M.sg**) followed (them.**M.pl**) wondered about the location of the hotel.	سائقي الحافلات الذين تبعهم الطفل تساءل عن موقع الفندق.
Double mismatch: (number + gender)	The bus drivers (**F.pl**) who the kid (**M.sg**) followed (them.**F.pl**) wondered about the location of the hotel.	سائقات الحافلات اللاتي تبعهن الطفل تساءل عن موقع الفندق.

Figure 5.1: Sample ORC stimuli with grammatical match conditions and English glosses. The matrix clause subject and resumptive object pronoun (which agree grammatically) are in yellow, and the relative clause noun and verb (which agree grammatically) are in blue. The disambiguating region – the relative clause verb – is circled in red.

For our analyses, we first analyzed the accuracy rates for the comprehension questions across clause types and grammatical conditions to determine whether increasing the number of grammatical cues in favor of an ORC interpretation significantly decreased the likelihood of a good-enough or noisy interpretation. If stronger grammatical cues mediate expectations and the acceptance of noisy interpretations, the results would show increased performance in comprehension question answers – and thus, less acceptance of a noisy interpretation – as the mismatch signals get stronger. Specifically, the highest level of noisy interpretations would be in the match condition, as the relative clause verb agrees with both the matrix and relative clause noun, and the object pronoun is the only signal of an ORC interpretation. There would then be less noisy interpretations in the single mismatch condition, where readers encounter two signals at the relative clause verb that they should be arriving at an ORC interpretation. Finally, the least amount of accepted noisy interpretations would be in the double mismatch condition, as this condition provides the highest number of signals of an ORC interpretation. However, if grammatical cues do not serve as a strong signal against noisy interpretations, the amount of noisy interpretations would be consistent despite stronger grammatical signals in the single and double mismatch conditions.

We then analyzed the eye movement data and tested whether we observed any significant interactions between grammatical conditions and clause type or correctness to see whether the different grammatical conditions had any effect on processing behaviors in ORCs, or on processing behaviors when arriving at the correct interpretation of an ORC versus an incorrect interpretation. Our main analyses focused on the same three comparative analyses from Experiment 3 (Table 4.1). We first investigated processing difficulty measures during veridical processing in all conditions to determine whether creating mismatch conditions affected processing difficulty in general. We focused on trials with correct comprehension question answers for SRCs and ORCs and determined whether a change in grammatical number or a change in grammatical gender created increased processing difficulty compared to the match condition. We then compared trials with misinterpreted ORCs to trials with correct ORCs and correct SRCs to

diagnose good-enough/noisy-channel processing, and whether these mismatch conditions also changed participants' processing behavior. In particular, we were interested in whether there was a direct behavioral difference in ORCs across different match conditions, and whether increasing the number of grammatical cues in favor of an ORC interpretation affected the processing behavior at all.

5.2. Methods

5.2.1. Participants

Thirty-one native Arabic speakers proficient in MSA (11 women, 20 men; mean age: 24; *SD* = 4.75) were recruited from the University of California, Davis (UCD). Participants were offered $30 in cash compensation for their participation, and some participants additionally received course credit. Prior to beginning the experiment, participants completed the same language history questionnaire from the first eye tracking experiment, in which they rated their proficiency in listening, speaking, reading and writing (scale from 1 to 7) in MSA and all their other known languages (see Appendix G for the full list of questions). Participants were considered proficient in MSA if they (1) selected Arabic as their native language, and (2) scored their proficiency in each area for MSA at 4 or higher. Three participants were excluded for rating one of their areas of proficiency in MSA lower than 4 (on a scale from 1 to 7, where 1 = low proficiency and 7 = high proficiency). As in our previous experiment, we established an a priori criterion to exclude any participant who scored lower than 75% accuracy on comprehension questions on filler items, but all participants performed above this criterion.

Since this participant population was more likely to include heritage speakers, we wanted to investigate some of the demographic information shared by our participants. Of the 31 participants whose data we used for our analyses, 29 (93.5%) indicated that they were born outside of the US in an Arabic-speaking country. 23 (74.2%) indicated that they also grew up in an Arabic-speaking country, while the remaining 8 indicated that they grew up somewhere within the US. When listing their native language, 26 participants included their specific dialect of Arabic, which represented 10 different dialects: 8 Egyptian,

5 Saudi, 3 Lebanese, 2 Palestinian, 2 Yemeni, 2 Iraqi, 1 Kuwaiti, 1 Algerian, 1 Syrian, and 1 Jordanian. The average Arabic proficiency (scale from 1 to 7) was 6.8 for listening, 6.7 for speaking, 6.5 for reading, and 6.4 for writing. When asked about how frequently they used Arabic (fixed options: Always, Often, Sometimes, Rarely, Never), 21 answered Always, 8 answered Often, and 2 answered Sometimes.

5.2.2. Materials

We used the same 80 items from our first eye tracking study, then created three different versions of each item with a different grammatical condition: match, single mismatch (number), and double mismatch (number and gender) (Figure 5.1). These grammatical manipulations were created for both SRCs and ORCs, resulting in six different conditions for each item: SRC match, ORC match, SRC mismatch, ORC mismatch, SRC double mismatch, and ORC double mismatch.

Our research question is interested in identifying whether specific types of cues, be they grammatical or orthographic, may have stronger effects in influencing the probability of a noisy interpretation. We are specifically interested in three types of differences that appear in our stimuli: differences by grammatical number, differences by grammatical gender, and differences by number of orthographic characters. It is important to note that our mismatch conditions do not have a one-to-one correspondence of differences by number, gender, and orthography: for example, there is a difference in grammatical number between the match condition and *both* mismatch conditions. Table 5.1 shows sample SRC and ORC stimuli in each grammatical condition and how the number, gender, and orthographic markings change in each region for each grammatical condition. When comparing grammatical and orthographic markings in the relative clause verb and RP clitic regions, there are three main contrasts that stand out. First, the main cue difference by grammatical number is between the match condition and both mismatch conditions. Second, the main cue difference by grammatical gender is between the single and double mismatch conditions, not directly between the match and double mismatch conditions. Finally, the only cue difference by number of orthographic characters occurs between the match condition and both mismatch conditions for items where the match condition is grammatically masculine. Crucially, masculine singular clitics are the only

clitics that are one character in length, while all other clitics are two characters long. We further discuss these differences and how we address them in our analyses in Section 5.3.3.

Table 5.1: Sample SRC and ORC stimuli in each grammatical condition annotated for their grammatical and orthographic markings. One example is provided for each grammatical gender in the match (i.e., baseline) condition: masculine (M) and feminine (F). The gender listed under the "manipulation" column indicates the grammatical gender of the matrix subject.

RC NP	**RP clitic**	**RC verb**	**RC pron**	**matrix NP**	**manipulation**
the kid (*M.sg*)	∅	followed (*M.sg*)	who (*M.sg*)	The bus driver (*M.sg*)	SRC-match-M
the kid (*M.sg*)	him (*M.sg; 1 chr*)	followed (*M.sg*)	who (*M.sg*)	The bus driver (*M.sg*)	ORC-match-M
the kid (*M.sg*)	∅	followed (*M.pl*)	who (*M.pl*)	The bus drivers (*M.pl*)	SRC-mismatch-M
the kid (*M.sg*)	them *(M.pl; 2 chr)*	followed (*M.sg*)	who (*M.pl*)	The bus drivers (*M.pl*)	ORC-mismatch-M
the kid (*M.sg*)	∅	followed (*F.pl*)	who (*F.pl*)	The bus drivers (*F.pl*)	SRC-double-mismatch-F
the kid (*M.sg*)	them *(F.pl; 2 chr)*	followed (*M.sg*)	who (*F.pl*)	The bus drivers (*F.pl*)	ORC-double-mismatch-F
the mother (*F.sg*)	∅	woke (*F.sg*)	who (*F.sg*)	The girl (*F.sg*)	SRC-match-F
the mother (*F.sg*)	her (*F.sg; 2 chr*)	woke (*F.sg*)	who (*F.sg*)	The girl (*F.sg*)	ORC-match-F
the mother (*F.sg*)	∅	woke (*F.pl*)	who (*F.pl*)	The girls (*F.pl*)	SRC-mismatch-F
the mother (*F.sg*)	them (*F.pl; 2 chr*)	woke (*F.sg*)	who (*F.pl*)	The girls (*F.pl*)	ORC-mismatch-F
the mother (*F.sg*)	∅	woke (*M.pl*)	who (*M.pl*)	The boys (*M.pl*)	SRC-double-mismatch-M
the mother (*F.sg*)	them (*M.pl; 2 chr*)	woke (*F.sg*)	who (*M.pl*)	The boys (*M.pl*)	ORC-double-mismatch-M

A norming study was conducted on our items in each condition to check that the changing of the grammatical number and gender did not affect the plausibility of the sentence. Native Arabic speakers (N = 120; 52 women, 66 men, 2 not reported; mean age: 30; SD = 9.53) were recruited through Prolific and asked to rate the plausibility of each sentence on a Likert scale (1 = highly implausible, 7 = highly

plausible). Participants for this norming study did not overlap with any participants from previous norming studies. Plausibility ratings were collected for both the full stimuli sentences and the relative clauses as simplified transitive sentences. 17 stimuli were excluded after an ANOVA revealed substantial discrepancies between plausibility ratings between the six clause type and grammatical conditions for those items. Two of the items received low ratings due to issues with translation, while the others appeared to be due to introducing a gender mismatch, which led to substantial plausibility shifts (e.g., female plumbers helping a male electrician was viewed as significantly less probable than male plumbers helping a male electrician). The mean plausibility rating for full stimuli sentences and simplified transitive sentences for each condition are reported in Table 5.2.

Table 5.2: Mean plausibility ratings and *SD*s for full stimuli sentences and simplified transitive sentences for each clause and grammatical match condition.

	Full sentences		**Simplified sentences**	
Stimuli type	***M***	***SD***	***M***	***SD***
ORCs				
match	6.06	1.47	5.94	1.53
single mismatch	6.00	1.50	6.03	1.44
double mismatch	5.99	1.53	5.98	1.48
SRCs				
match	5.97	1.58	6.10	1.42
single mismatch	5.98	1.50	6.09	1.37
double mismatch	5.89	1.54	5.95	1.49

Exclusions were made after data collection; participants saw all 80 stimuli plus 80 unrelated filler items for a total of 160 items. Exclusions from the norming study resulted in 63 total stimuli for our analyses. Experimental items were split into six lists and counterbalanced within and across lists in a Latin square design. Participants were randomly assigned to one of six lists, and the order of sentences within each list was randomized separately for each participant. Comprehension questions targeting comprehension of the relative clause appeared after all stimuli, following the design of our previous experiment, and general comprehension questions appeared after all filler items. We randomized whether the correct or incorrect answer was presented first and counterbalanced by item type.

5.2.3. Apparatus and Procedure

The experiment took place at UCD in the Department of Linguistics' eye tracking lab using the Eyelink 1000 Plus eye tracker (SR Research). Right eye gaze movements were recorded via a high-speed 35 mm lens on a tower mount at a sampling rate of 1000 Hz. Participants' head movements were stabilized using a head stabilization tower. Sentences were displayed on a high definition (1920 x 1080 pixels) 24" BENQ ZOWIE XL 2430 monitor at 80 cm viewing distance. Text was presented in 20-point Times New Roman, a proportionally spaced typeface.

Participants were tested individually in an isolated, sound-proof room. Instructions were given verbally in English and were presented in both English and MSA text on the screen. The eye tracker was calibrated and validated using the horizontal 3-point calibration. Participants completed two practice trials and were allowed to ask questions before proceeding with the experiment.

The experiment was divided into eight blocks of 20 items each to allow for breaks throughout the experiment. Calibration accuracy was assessed before the start of each block and the eye tracker was re-calibrated as necessary. At the beginning of each trial, a right-aligned asterisk was placed at the onset site of the first letter of the sentence. Once the participant fixated for at least 800 ms, the sentence appeared and replaced the asterisk. Participants pressed the spacebar on the keyboard once they were finished reading the sentence, then used the mouse to click and select the correct comprehension question answer on the following screen. Completion of the experiment took about 60 minutes for each participant.

5.3. Analysis and Results

Data were cleaned using SR Research's Data Viewer. Following standard procedures, fixations that were less than 80 ms and within one character of each other were merged, and remaining fixations less than 80

ms or longer than 1,200 ms were excluded. We also excluded trials where significant track loss occurred and fixations where a blink occurred on the target word. This resulted in a total data loss of 20.6%[6].

Interest areas were divided into the same regions as in Experiment 3. For each region, we calculated the same eye tracking metrics using the Get Reading Measures package from SR Research: *first fixation duration*, *first pass duration*, *go-past time*, *total fixation duration*, *first pass regression*, and *first pass skip*. In terms of number of trials, we observed 888 correct SRC trials, 717 correct ORC trials, and 233 misinterpreted ORC trials. Notably, the number of trials for these analyses are lower than those from Experiment 3, so we keep this in mind in the interpretation of our results.

5.3.1. Comprehension question analysis

Our first research question was whether additional grammatical cues in favor of an ORC interpretation and against a noisy SRC interpretation had a significant effect on the number of misinterpreted ORCs. Specifically, we were interested in exploring whether trials in the match condition had significantly more incorrect interpretations of ORCs compared to trials in the single or double mismatch condition.

We first calculated some summary statistics about accuracy rates and found an overall accuracy rate of 84.7%, with a 93.7% accuracy for SRCs compared to only 75.5% accuracy for ORCs. These accuracy rates aligned very closely with the accuracy rates in Experiment 3. We also calculated accuracy by participant and once again found substantial variance by individual. ORC scores ranged from 18% to 97%; six participants (out of 31) scored less than 50%, while no participant scored 100%. Conversely, the lowest score on SRCs was 67%, with the second lowest score being 79%, and eight participants scored 100%. A distribution of the average scores by participant and clause type are shown in Figure 5.2.

[6] This data loss percentage is substantially higher than in our previous experiments, where data loss due to data cleaning was 5-6%. We attribute this higher data loss to be due to a larger number of logistical issues (e.g., eye tracker calibration errors, participant distractibility, etc.) than in previous experiments.

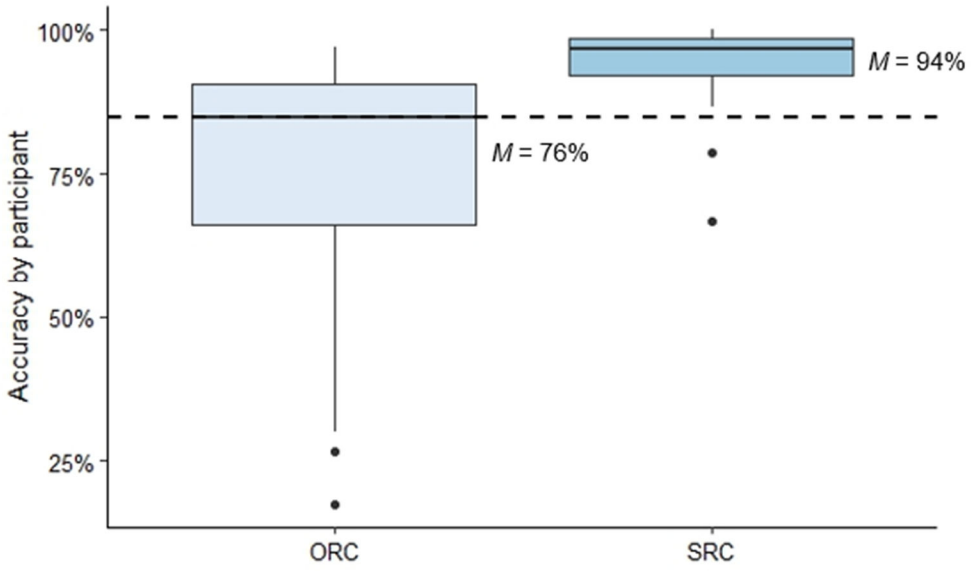

Figure 5.2: Box plot of average accuracy by participant for each clause type. The horizontal dashed line represents the overall mean of 85%.

We then calculated accuracy rates by grammatical condition within each clause type. Mean accuracy rates by condition and clause type showed that SRCs had consistent accuracy rates across conditions, but ORCs differed across conditions (Table 5.3).

Table 5.3: Trial accuracy rates and SDs by clause type and grammatical match condition.

	Match		**Single mismatch**		**Double mismatch**	
Stimuli type	***M***	***SD***	***M***	***SD***	***M***	***SD***
ORCs	75.5%	0.43	79.1%	0.41	71.9%	0.45
SRCs	93.1%	0.25	94.2%	0.23	94.3%	0.23

To test the statistical significance of these differences, we fit a logistic mixed-effects regression model to the data with Correctness as the dependent variable and Clause Type (ORC: 1, SRC: -1) and Grammatical Condition (Match: 1, 0; Single Mismatch: -1, -1; Double Mismatch: 0, 1) as sum-coded fixed effects, plus their interaction. We also used the maximal random effects structure by Participant and Item. Model estimates for main effects and interactions are reported in Table 5.4.

Table 5.4: Logistic mixed-effects model estimates of each dependent variable and interaction on correct comprehension question answers, including SE estimates and CrIs. Estimates marked with an asterisk are significant.

Main Effects	**β**	***SE***	**CrI**
Clause Type	1.04*	0.17	[0.71, 1.38]
Grammatical Condition (Match)	-0.19	0.16	[-0.51, 0.13]
Grammatical Condition (Double Mismatch)	-0.07	0.18	[-0.41, 0.29]
*Clause Type * Grammatical Condition (Match)*	-0.12	0.16	[-0.43, 0.20]
*Clause Type * Grammatical Condition (Double Mismatch)*	0.16	0.16	[-0.15, 0.48]

Model estimates showed a significant effect of Clause Type on correctness, such that ORCs were more likely to receive incorrect answers than SRCs. Yet, there were no significant effects for Grammatical Condition on correctness, nor any significant interactions between Clause Type and Grammatical Condition. These results demonstrate that clause type did have a significant effect on whether a participant arrived at the correct interpretation of the sentence, which aligns with our findings from Experiment 3. However, the number of cues provided in favor of a particular interpretation – as represented by the different grammatical conditions – did not significantly affect the likelihood of a correct interpretation. Despite receiving additional grammatical cues in favor of an ORC interpretation and against an SRC interpretation, readers still consistently accepted good-enough or noisy SRC interpretations of ORCs.

5.3.2. Investigating significant interactions

Our second research question asked whether providing additional grammatical cues in favor of an ORC interpretation and against a noisy SRC interpretation affected participants' reading behaviors, particularly when reading ORCs. Consequently, one of the main effects of interest in our analyses was whether we observed significant interactions between clause type (SRC or ORC) or correctness (whether an ORC was correctly interpreted) and grammatical condition. To investigate this possibility, we conducted an omnibus test (Doornik & Hansen, 2008).

In addition to investigating these interactions, we also tested whether including another control predictor was a helpful addition to our models. One potential confound that arose as we were preparing our analyses was whether our single and double mismatch conditions could be conflated with differences in gender. The original items used in Experiment 3 were balanced for grammatical gender; however, the numerous exclusions from the norming study in Experiment 4 resulted in an imbalance of grammatical gender across stimuli. Of the 63 total items, 36 items had grammatically masculine matrix subjects in the match and single mismatch condition, while only 27 had grammatically feminine matrix subjects. These numbers then swapped for the double mismatch condition, as this condition created a mismatch in both number *and* gender: 36 items in the double mismatch condition had grammatically feminine matrix subjects and 27 had grammatically masculine matrix subjects. We thus decided to also include grammatical gender as a fixed effect as part of our omnibus analysis.

5.3.2.1. Effect of item type with all correct trials

We first focused on whether there were significant interactions between clause type and grammatical condition within all of our items with correct comprehension questions (i.e., Cor ORCs vs. Cor SRCs). We focused only on our primary regions of interest – the relative clause verb and relative clause NP – and on our primary fixation measures of interest – first pass times, go-past times, total fixation durations, and first pass regressions. We started by fitting models that included sum-coded fixed effects for Clause Type (ORC: 1, SRC: -1), Grammatical Condition (Match: 1, 0; Single Mismatch: -1, -1; Double Mismatch: 0, 1), and Matrix Subject Gender (F: 1, M: -1), plus all possible two-way interactions[7]. We also included control predictors of Word Length (numeric) and Trial Index (numeric) and used the maximal random effect structure by Participant and Item.

Model estimates revealed significant effects for all three main effects, and some significant interactions between Grammatical Condition and Matrix Subject Gender. However, there were no

[7] We also considered a three-way interaction with matrix subject gender, plus all other possible two-way interactions, but no models that included interactions with matrix subject gender survived WAIC comparison.

significant interactions between Clause Type and Grammatical Condition, or Clause Type and Matrix Subject Gender. This finding was surprising, as we expected that changing the grammatical condition of an item would have a stronger effect on ORCs than SRCs due to the relative clause verb and resumptive object pronoun mismatch. However, this result was consistent with the findings from our comprehension question analysis, which revealed that grammatical condition had no significant effect on the accuracy of a trial. Since interactions between Clause Type and Grammatical Condition and Clause Type and Matrix Subject Gender were not significant, we removed these from our model considerations.

To further investigate the significant interaction between Grammatical Condition and Matrix Subject Gender, we fit two new sets of models: one that included Clause Type, Grammatical Condition, Matrix Subject Gender, and a two-way interaction between Grammatical Condition and Matrix Subject Gender, and one that included all these main effects with no interactions. We then conducted a WAIC analysis to determine which model architecture best fit our data (Watanabe, 2010). WAIC estimates are reported in Table 5.5. The WAIC analysis showed that both sets of models did equally well for some measures, but the purely additive model outperformed the model with the two-way interaction in the remaining measures. Since we find no strong evidence of a significant interaction effect with Matrix Subject Gender, and because it is not a primary variable of interest, we settled on the purely additive model.

Table 5.5: WAIC analysis estimates for Cor ORC vs. Cor SRC models. The interaction models included a two-way interaction between Grammatical Condition and Matrix Subject Gender, and the additive models did not include any interactions. An *elpd_diff* score of 0 indicates the better fitting model for the data. Scores that are significantly different are marked with an asterisk.

Fixation Metrics	**Interaction model**		**Additive model**	
by Region	**elpd_diff**	***SE***	**elpd_diff**	***SE***
RC verb				
first pass	-1.29	6.03	0.00	0.00
go-past	0.00	0.00	-0.59	6.00
total duration	-0.55	4.90	0.00	0.00
p(regress)	-4.33	2.23	0.00*	0.00
RC NP				
first pass	-2.40	6.02	0.00	0.00
go-past	0.00	0.00	-5.19	9.17
total duration	-3.06	2.27	0.00*	0.00
p(regress)	-3.24	2.92	0.00*	0.00

As a final check of a possible trend of an interaction between clause type and grammatical condition, we plotted average reading times by clause type and condition to see if we observed any numerical differences in reading times across conditions. Average first pass, go-past and total fixation duration times in the relative clause verb are plotted in Figure 5.3, and the same metrics in the relative clause NP region are plotted in Figure 5.4. Is important to note that average total fixation duration times include by definition first pass and go-past times, and go-past times include first pass times.

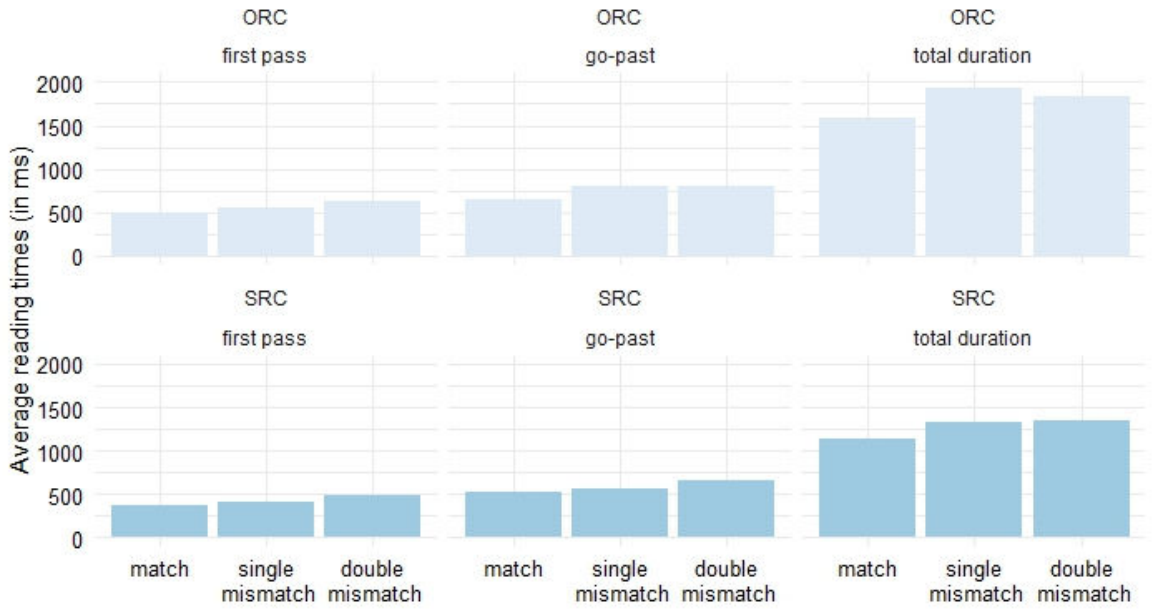

Figure 5.3: Average first pass, go-past and total fixation duration times in the relative clause verb region for all correct trials, by Clause Type and Grammatical Condition.

At the relative clause verb region, SRC trials showed a slight increase in reading times across grammatical conditions, with items in the match condition being read faster than items in the mismatch conditions, and items in the single mismatch condition being read slightly faster than items in the double mismatch condition. ORC trials appeared to show a similar trend for first pass and go-past reading times but had the longest average total fixation durations for the single mismatch condition. Notably, first pass

and go-past reading times are indicative of earlier processing mechanisms, while total fixation duration is indicative of later processing mechanisms.

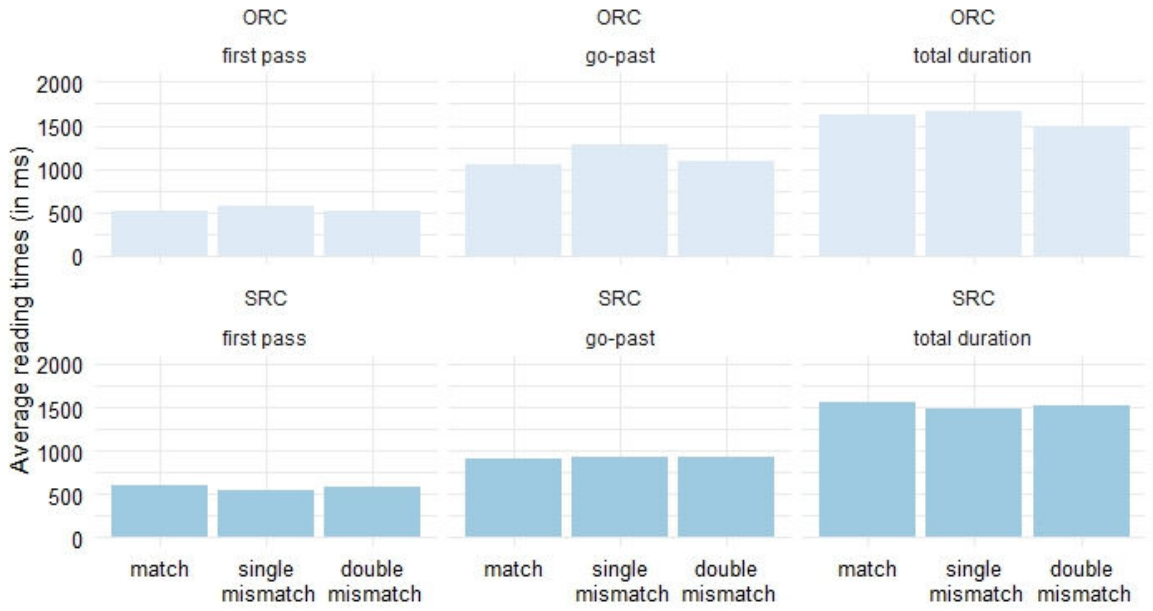

Figure 5.4: Average first pass, go-past and total fixation duration times in the relative clause NP region for all correct trials, by Clause Type and Grammatical Condition

At the relative clause NP, SRC trials showed no real discernible pattern in differences between grammatical conditions. On the other hand, ORC trials had longer average first pass and go-past times for the single mismatch condition compared to the match and double mismatch conditions. Total fixation duration times in this region were longer for the match and single mismatch conditions, but shorter for the double mismatch conditions.

We had predicted that reading times would be the fastest in the match condition and would be the slowest in the double mismatch condition, entailing that adding additional cues in favor of an ORC interpretation would increase processing cost and slow down reading times proportional to the number of cues added. However, these data suggest that the biggest slowdown happens due to a change in grammatical number, from singular in the match condition to plural in the single mismatch condition, and

that adding the additional cue of gender has the reverse effect. This perhaps suggests that grammatical number is a more important cue than grammatical gender when identifying grammatical roles of arguments within a sentence. However, since these analyses are based on a small number of trials, more data must be collected in order to provide more conclusive evidence for this speculation.

5.3.2.2. Effect of correctness with all ORC trials

We then wanted to investigate any possible interactions of Grammatical Condition with Correctness, as this variable was distinct from clause type and our previous analysis did not consider effects within incorrect trials. To do this, we focused on all ORC trials and compared correctly interpreted trials to misinterpreted trials (i.e., Incor ORCs vs. Cor ORCs). We again focused only on our regions and fixation measures of interest. The first set of models included sum-coded fixed effects for Correctness (Correct: -1, Incorrect: 1), Grammatical Condition (Match: 1, 0; Single Mismatch: -1, -1; Double Mismatch: 0, 1), and Matrix Subject Gender (F: 1, M: -1), plus a two-way interaction between Correctness and Grammatical Condition. We did not consider any further interactions with Matrix Subject Gender as the previous analysis showed that these interactions were not significant, and the analyses considering misinterpreted ORC trials have a smaller number of trials than the vertical processing analysis, making it even less likely that we would be able to reliably detect interaction effects. These models also included control predictors of Word Length (numeric) and Trial Index (numeric) and used the maximal random effect structure by Participant and Item. The second set of models included all of the same fixed and random effects but did not include the two-way interaction between Correctness and Grammatical Condition.

Following the first omnibus analysis, we conducted a WAIC analysis to determine which model architecture best fit our data. WAIC estimates are reported in Table 5.6. Similar to the first analysis, the WAIC analysis showed that both sets of models did equally well for some measures, but the purely additive model again outperformed the model with the two-way interaction in the remaining measures. Thus, even when considering processing differences between correctly interpreted and misinterpreted

ORCs, grammatical condition does not significantly affect the reading behavior associated with a correct interpretation versus an incorrect interpretation.

Table 5.6: WAIC analysis estimates for Incor ORC vs. Cor ORC models. The interaction models included a two-way interaction between Correctness and Grammatical Condition, and the additive models did not include any interactions. An *elpd_diff* score of 0 indicates the better fitting model for the data. Scores that are significantly different are marked with an asterisk.

Fixation Metrics	**Interaction model**		**Additive model**	
by Region	**elpd_diff**	***SE***	**elpd_diff**	***SE***
RC verb				
first pass	0.00	0.00	-19.02	20.25
go-past	-0.27	7.93	0.00	0.00
total duration	-7.06	1.61	0.00*	0.00
p(regress)	-6.75	2.11	0.00*	0.00
RC NP				
first pass	-3.04	4.24	0.00	0.00
go-past	-3.32	2.71	0.00*	0.00
total duration	-4.75	2.65	0.00*	0.00
p(regress)	-3.36	3.00	0.00*	0.00

5.3.2.3. Effect of item type on misinterpreted ORC trials versus correct SRC trials

Finally, we wanted to test whether we saw any significant interactions specifically when comparing misinterpreted ORC trials to correct SRC trials (i.e., Incor ORCs vs. Cor SRCs). We fit the same types of models as for the Incor ORCs vs. Cor ORCs omnibus analysis, but with Clause Type (ORC: 1, SRC: -1) instead of Correctness. WAIC estimates comparing the two types of models are reported in Table 5.7. The estimates showed that in two instances, the interaction model was the best fit, and in two other instances, the additive model was the best fit.

To explore these few significant interactions, we again plotted average reading times by Clause Type and Grammatical Condition to see if we observed any numerical differences in reading times across conditions. We focused on the two fixation metrics for which the interaction model was a better fit: first pass and go-past times in the relative clause verb region. These averages are plotted in Figure 5.5.

Table 5.7: WAIC analysis estimates for Incor ORC vs. Cor SRC models. The interaction models included a two-way interaction between Item Type and Grammatical Condition, and the additive models did not include any interactions. An *elpd_diff* score of 0 indicates the better fitting model for the data. Scores that are significantly different are marked with an asterisk.

Fixation Metrics	**Interaction model**		**Additive model**	
by Region	**elpd_diff**	***SE***	**elpd_diff**	***SE***
RC verb				
first pass	0.00*	0.00	-33.86	23.39
go-past	0.00*	0.00	-34.21	21.68
total duration	0.00	0.00	-0.63	4.69
p(regress)	-4.03	3.00	0.00*	0.00
RC NP				
first pass	-2.00	6.11	0.00	0.00
go-past	-2.07	3.83	0.00	0.00
total duration	-3.52	2.94	0.00*	0.00
p(regress)	-2.30	3.30	0.00	0.00

Reading time averages for correct SRC trials are the same as shown in Figure 5.3 in Section 835.3.3.1, and showed a slight increase in reading times across grammatical conditions. Reading time averages for misinterpreted ORC trials showed a more pronounced difference between grammatical conditions, with the longest reading times in the Single Mismatch condition compared to the Match or Double Mismatch conditions. These averages do not align with the averages for correctly interpreted ORCs (Figure 5.3 in Section 5.3.3.1), which showed longer fixation times for the Double Mismatch condition for these two metrics at the relative clause verb. However, this pattern *does* match the pattern for these metrics at the relative clause NP (Figure 5.4 in Section 5.3.2.1): the Single Mismatch condition had the longest first pass and go-past times for ORCs in that region. Despite these numeric differences, we still find no significant interactions between Correctness and Grammatical Condition when comparing misinterpreted ORCs and correct ORCs, and the number of trials for misinterpreted ORCs (233 trials) compared to correct SRCs (888 trials) was more imbalanced in this analysis. Since the additive model was the best fit for the two previous comparisons, we continue to use a purely additive model in our future analyses for this comparison as well.

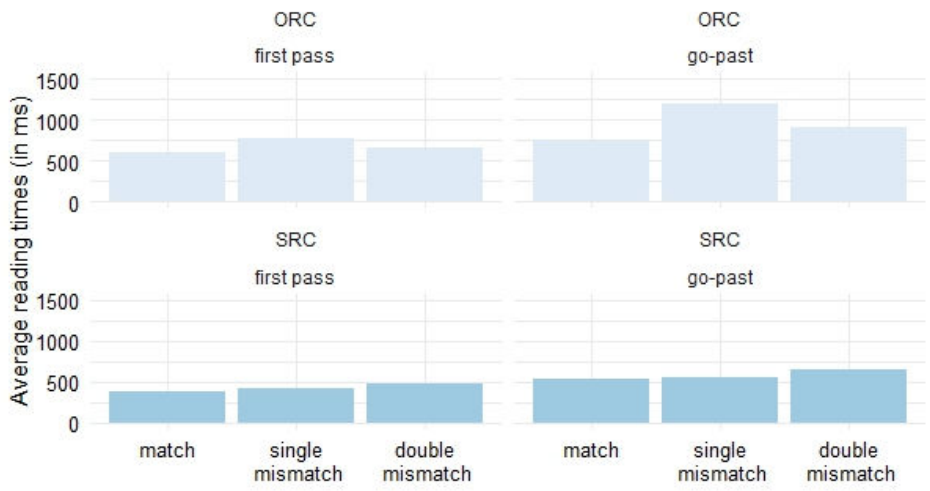

Figure 5.5: Average first pass and go-past times in the relative clause verb region for misinterpreted ORCs and correct SRCs, by clause type and grammatical condition.

Taken together, these three analyses demonstrate that our data across the board are best modeled without considering any interactions between main effects. Thus, our data appear to indicate that grammatical condition does not have a significant effect on ORCs compared to SRCs, nor does it have a significant effect on the likelihood of a good-enough or noisy interpretation. These findings are tentative and will require more data to be conclusive: since we had far fewer trials for Experiment 4 than Experiment 3, and further investigated six different conditions compared to just two, it is possible that we do not have the statistical power necessary to detect these significant interaction terms. We hope to collect more data in the future to further investigate this possibility.

5.3.3. Main analyses

Finally, we revisited our main analyses as conducted in Experiment 3 (Table 4.1). Since our omnibus analyses ruled out any significant interactions between grammatical condition and our other primary variables of interest, these analyses focused on testing the main effects of item type and grammatical condition on reading behaviors. Specifically, we aimed to compare reading behaviors from this experiment to those in Experiment 3, and test whether we again see support for good-enough or noisy-

channel processing over misreading. We also considered whether grammatical condition had any significant main effects on reading behaviors.

As with previous analyses, we fit linear mixed-effects regression models for numeric reading measures and logistic mixed-effects regression models for binary measures, and fit individual models for each eye tracking metric in each interest area for each analysis.

In terms of coding our main effects, we took a different approach from our typical sum coding in order to provide more interpretable regression results. We were interested in investigating two different effects: first, whether there was a significant difference between the match condition and *either* mismatch condition; and second, whether there was a significant difference between the two mismatch conditions. Coding for these differences, rather than simply the differences between the various grammatical conditions, allows us to isolate significant differences by grammatical number and gender cues (see Section 5.2.2). To estimate these two effects, we included treatment-coded fixed effects for Clause Type (ORC: 1, SRC: 0 for Cor ORC vs. Cor SRC and Incor ORC vs. Cor SRC models) or Correctness (Correct: 0, Incorrect: 1 for Incor ORC vs. Cor ORC models), Matrix Subject Gender (M: 0, F: 1), and a custom treatment coding for Grammatical Condition (Match: 0, 0; Single Mismatch: 1, -0.5; Double Mismatch: 1, 0.5). This custom treatment coding resulted in one estimate for the difference between the match condition and either mismatch condition, and a second estimate for the difference between the two mismatch conditions above and beyond the overall difference between the match and mismatch conditions. Identical to our previous analyses, we also included control predictors of Word Length (numeric) and Trial Index (numeric) and used the maximal random effect structure by Participant and Item. Model estimates for main effects in the regions of interest – the relative clause verb and NP – are reported in Table 5.8 and Table 5.9, and model estimates for all other regions are reported in Appendix H.

Table 5.8: Linear and logistic mixed-effects model estimates of the dependent variables on all fixation metrics in the relative clause verb and NP regions for the Cor ORC vs. Cor SRC analysis, including SE estimates and CrIs. The % > 0 column shows the percent of the sampled posterior distribution that is over or under 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates that are bolded are significant.

	Cor ORC vs. Cor SRC							
	RC verb				**RC NP**			
Fixation Metrics	**β**	***SE***	**CrI**	**% >0**	**β**	***SE***	**CrI**	**% >0**
first fixation								
ORC	7.38	11.21	[-14.85, 29.20]	75	-9.93	12.77	[-34.57, 15.99]	21
Mismatch vs. Match	5.63	9.77	[-13.47, 24.79]	72	-10.93	8.84	[-28.35, 6.37]	11
Double vs. Single	15.26	12.66	[-9.49, 40.54]	89	-11.49	11.04	[-33.05, 10.31]	15
Feminine	10.53	10.30	[-9.92, 30.83]	85	2.34	12.20	[-21.64, 26.68]	57
first pass								
ORC	**102.59**	**35.77**	**[33.75, 173.82]**	**100**	-17.04	26.85	[-70.65, 34.54]	27
Mismatch vs. Match	26.28	26.36	[-24.90, 79.21]	84	-57.56	43.89	[-146.33, 27.31]	9
Double vs. Single	**59.34**	**27.44**	**[5.27, 113.50]**	**98**	12.53	40.53	[-65.75, 93.70]	62
Feminine	**74.26**	**29.53**	**[15.82, 131.87]**	**99**	-32.88	30.29	[-92.79, 27.34]	14
go-past								
ORC	**114.49**	**42.77**	**[31.39, 199.11]**	**100**	**204.92**	**75.94**	**[57.17, 352.10]**	**100**
Mismatch vs. Match	59.90	38.23	[-14.40, 135.09]	94	31.35	64.34	[-95.78, 158.38]	69
Double vs. Single	57.64	43.47	[-26.45, 143.75]	91	-85.34	96.82	[-275.92, 107.47]	18
Feminine	**111.17**	**41.10**	**[29.53, 191.74]**	**100**	**109.96**	**68.46**	**[-24.88, 243.64]**	**95**
total duration								
ORC	**392.59**	**77.93**	**[238.74, 546.10]**	**100**	55.10	87.64	[-119.24, 227.28]	74
Mismatch vs. Match	**159.29**	**72.99**	**[15.67, 301.76]**	**98**	-90.78	60.58	[-209.61, 27.20]	7
Double vs. Single	-55.30	79.68	[-211.13, 101.50]	24	-58.84	80.31	[-216.08, 97.24]	23
Feminine	**341.12**	**80.00**	**[184.07, 499.75]**	**100**	24.24	67.27	[-108.59, 155.40]	64
p(regress)								
ORC	0.13	0.18	[-0.22, 0.48]	78	**0.38**	**0.14**	**[0.10, 0.66]**	**100**
Mismatch vs. Match	0.07	0.18	[-0.29, 0.43]	66	0.16	0.16	[-0.15, 0.47]	85
Double vs. Single	0.03	0.24	[-0.45, 0.51]	54	-0.10	0.18	[-0.46, 0.25]	28
Feminine	0.04	0.20	[-0.36, 0.41]	59	0.17	0.15	[-0.12, 0.45]	88
p(skip)								
ORC	0.16	0.27	[-0.40, 0.66]	74	0.02	0.29	[-0.57, 0.58]	54
Mismatch vs. Match	-0.03	0.26	[-0.54, 0.47]	45	-0.01	0.35	[-0.76, 0.64]	52
Double vs. Single	-0.71	0.35	[-1.42, -0.06]	2	-0.08	0.40	[-0.93, 0.66]	43
Feminine	-0.17	0.30	[-0.79, 0.40]	29	-0.20	0.36	[-0.97, 0.45]	30

Table 5.9: Linear and logistic mixed-effects model estimates of the dependent variables on all fixation metrics in the relative clause verb and NP regions for the Incor ORC vs. Cor ORC and Incor ORC vs. Cor SRC analyses, including SE estimates and CrIs. The main effect for the Incor ORC vs. Cor ORC analysis is "Incorrect" while the main effect for the Incor ORC vs. Cor SRC analysis is "ORC," which is indicated by "Incorrect | ORC." The % > 0 column shows the percent of the sampled posterior distribution that is over or under 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates that are bolded are significant.

	Incor ORC vs. Cor ORC								**Incor ORC vs. Cor SRC**							
	RC verb				**RC NP**				**RC verb**				**RC NP**			
Fixation Metrics	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**
first fixation																
Incorrect | ORC	12.47	18.28	[-23.27, 48.76]	75	8.82	15.89	[-22.60, 40.68]	71	21.19	18.45	[-15.51, 56.84]	88	6.40	13.79	[-20.64, 33.14]	68
Mismatch vs. Match	6.52	13.33	[-19.56, 32.47]	69	-14.67	12.60	[-39.64, 10.16]	12	9.29	11.82	[-14.08, 32.49]	79	-8.22	10.86	[-29.61, 13.08]	23
Double vs. Single	0.99	15.05	[-28.78, 30.63]	52	-14.52	14.03	[-41.94, 13.13]	15	21.41	15.36	[-8.37, 51.91]	92	-11.75	13.26	[-37.97, 14.14]	19
Feminine	3.18	16.67	[-29.87, 35.36]	58	14.41	13.50	[-12.13, 41.08]	86	16.42	11.90	[-6.89, 40.12]	92	-1.50	13.09	[-27.49, 24.35]	45
first pass																
Incorrect | ORC	9.96	66.07	[-117.56, 141.57]	55	67.70	52.61	[-36.85, 173.60]	90	123.93	77.30	[-25.36, 274.99]	94	-26.72	47.35	[-118.46, 66.74]	29
Mismatch vs. Match	36.57	35.16	[-32.36, 105.18]	85	10.71	33.65	[-54.97, 76.49]	63	16.90	29.77	[-42.39, 74.78]	72	-42.25	44.60	[-129.98, 45.97]	17
Double vs. Single	30.56	46.81	[-63.01, 122.61]	75	7.79	44.48	[-80.66, 96.3]	57	14.25	28.74	[-42.05, 71.29]	69	31.05	49.51	[-66.87, 128.51]	73
Feminine	**74.57**	**41.37**	**[-6.73, 155.81]**	**96**	-55.11	37.76	[-129.51, 19.22]	7	**109.41**	**28.97**	**[52.71, 166.80]**	**100**	-38.48	36.16	[-109.61, 32.23]	14
go-past																
Incorrect | ORC	109.80	72.60	[-32.97, 255.12]	94	-106.74	109.04	[-319.42, 107.92]	16	**206.57**	**85.73**	**[39.13, 378.81]**	**99**	-36.59	77.63	[-190.12, 115.10]	32
Mismatch vs. Match	**121.99**	**53.42**	**[17.03, 227.64]**	**99**	37.86	91.78	[-141.97, 218.96]	66	51.40	44.52	[-34.97, 139.34]	88	-61.05	67.86	[-194.18, 73.78]	18
Double vs. Single	-38.09	81.83	[-201.36, 121.81]	32	-100.76	124.47	[-349.85, 144.07]	20	21.46	49.30	[-75.55, 119.37]	67	-14.69	81.67	[-173.97, 145.92]	43
Feminine	**140.61**	**63.17**	**[15.92, 265.17]**	**99**	**165.65**	**93.02**	**[-17.36, 347.36]**	**96**	**112.75**	**38.96**	**[36.49, 190.23]**	**100**	94.28	67.04	[-39.18, 223.69]	92
total duration																
Incorrect | ORC	-126.03	109.01	[-339.58, 89.52]	12	-4.94	98.04	[-198.87, 190.04]	48	**184.16**	**95.34**	**[0.63, 375.64]**	**98**	-33.07	101.22	[-225.62, 171.57]	36
Mismatch vs. Match	**173.19**	**103.35**	**[-31.87, 375.62]**	**95**	-23.90	85.13	[-189.92, 145.08]	39	42.48	76.49	[-108.88, 190.23]	71	-64.84	69.30	[-201.56, 70.60]	17
Double vs. Single	-149.65	107.12	[-362.41, 58.48]	8	-87.68	89.77	[-262.3, 89.98]	16	-55.67	85.61	[-223.27, 113.59]	25	34.99	89.61	[-140.83, 211.54]	66
Feminine	**334.37**	**104.91**	**[126.37, 540.52]**	**100**	16.67	72.14	[-123.76, 158.77]	59	**387.49**	**80.33**	**[228.30, 545.34]**	**100**	108.57	73.19	[-37.64, 250.45]	93
p(regress)																
Incorrect | ORC	0.30	0.28	[-0.26, 0.83]	86	**-0.61**	**0.28**	**[-1.19, -0.09]**	**1**	0.37	0.28	[-0.20, 0.91]	91	-0.24	0.27	[-0.82, 0.25]	18
Mismatch vs. Match	0.02	0.28	[-0.53, 0.57]	53	-0.01	0.19	[-0.39, 0.36]	47	0.14	0.22	[-0.30, 0.58]	73	-0.02	0.18	[-0.37, 0.33]	46
Double vs. Single	-0.37	0.29	[-0.94, 0.19]	10	-0.08	0.26	[-0.6, 0.41]	37	0.14	0.25	[-0.34, 0.65]	71	-0.09	0.22	[-0.51, 0.34]	33
Feminine	0.17	0.24	[-0.30, 0.63]	76	0.25	0.19	[-0.13, 0.62]	91	0.03	0.20	[-0.38, 0.41]	56	0.15	0.19	[-0.22, 0.51]	79

p(skip)															
Incorrect | ORC	**-1.17**	**0.77**	**[-3.04, -0.03]**	**2**	**-1.10**	**0.76**	**[-2.92, 0.06]**	**3**	-0.81	0.69	[-2.51, 0.23]	8	**-1.11**	**0.72**	**[-2.78, 0.03]**	**3**
Mismatch vs. Match	-0.32	0.38	[-1.11, 0.41]	20	0.03	0.45	[-0.93, 0.86]	55	0.15	0.32	[-0.47, 0.78]	69	-0.09	0.57	[-1.29, 0.99]	45
Double vs. Single	-0.86	0.51	[-1.94, 0.09]	4	0.03	0.47	[-0.94, 0.93]	54	-0.37	0.37	[-1.14, 0.35]	16	-0.06	0.58	[-1.26, 1.04]	47
Feminine	-0.31	0.43	[-1.20, 0.49]	23	-0.45	0.38	[-1.24, 0.24]	11	-0.17	0.33	[-0.86, 0.43]	31	-0.14	0.53	[-1.28, 0.80]	42

5.3.3.1. Correct ORCs versus Correct SRCs

We found significant effects for Clause Type at the relative clause verb for first pass, go-past, and total fixation duration, such that ORCs were read slower than SRCs in this region. We also found significant effects by Grammatical Condition: the single mismatch condition had faster first pass reading times and a higher probability of a first pass skip than the double mismatch condition, and both mismatch conditions had slower total fixation durations than the match condition.

We also found a significant effect for Clause Type for total fixation duration and first pass regressions at the relative clause NP, such that ORCs were also read slower than SRCs in this region and had a higher probability of a first pass regression. We did not find any significant effects by Grammatical Condition.

5.3.3.2. Misinterpreted ORCs versus Correct ORCs

We found a significant effect for Correctness at the relative clause verb for first pass skips, such that misinterpreted ORC trials were less likely to skip the relative clause verb region on pass. However, we found no significant effects for Correctness for any primary fixation metric in this region. We also found significant effects by Grammatical Condition: both mismatch conditions had longer go-past and total fixation durations than the match condition at the relative clause verb, and the single mismatch condition had a higher probability of a first pass skip at the relative clause verb than the double mismatch condition.

We also found significant effects for Correctness at the relative clause NP for first pass regressions and skips, such that misinterpreted ORC trials were less likely to make a first pass regression at the relative clause NP and less likely to skip it on first pass.

Like our analyses for Experiment 3, we also wanted to look at fixation metrics on the resumptive pronoun clitic specifically while comparing misinterpreted versus correct ORCs. To test the statistical significance of these differences, we ran additional regression models in this region following the same design outlined above. The model estimates for main effects are reported in Table 5.10. We found no significant effects for Correctness on any fixation metric; however, there were multiple significant

differences by Grammatical Condition and Matrix Subject Gender. Items in the mismatch conditions were read slower and were more likely to be skipped on first pass than items in the match condition. Items whose matrix subject was grammatically feminine were also read slower, but were less likely to be skipped on first pass. Notably, these effects are above and beyond any orthographic differences in the clitics as our models included a control parameter for Word Length (i.e., the length of the clitic); further, there were no significant differences by length alone.

Table 5.10: Linear and logistic mixed-effects model estimates of the dependent variables on all fixation metrics in the resumptive pronoun clitic region, including SE estimates and CrIs. Estimates marked with an asterisk are significant.

Fixation Metrics	**Incor ORC vs. Cor ORC**			
at RP clitic	**β**	***SE***	**CrI**	**% >0**
first fixation				
Incorrect	-4.41	14.11	[-31.83, 23.37]	38
Mismatch vs. Match	37.37*	16.47	[4.79, 69.38]	99
Single vs. Double	-2.49	13.13	[-28.10, 22.95]	43
Feminine	32.94*	12.98	[7.62, 58.62]	99
first pass				
Incorrect	-9.23	16.76	[-42.07, 23.83]	29
Mismatch vs. Match	51.49*	20.69	[10.46, 91.62]	99
Single vs. Double	-12.68	15.44	[-42.93, 17.36]	21
Feminine	43.08*	18.06	[7.06, 78.17]	99
go-past				
Incorrect	-23.22	37.31	[-94.78, 52.26]	26
Mismatch vs. Match	84.63*	42.88	[1.84, 168.83]	98
Single vs. Double	-41.14	33.77	[-106.43, 23.95]	12
Feminine	46.95	33.67	[-18.78, 113.78]	92
total duration				
Incorrect	-48.00	46.38	[-137.34, 45.32]	15
Mismatch vs. Match	176.05*	55.60	[64.34, 283.71]	100
Single vs. Double	17.99	47.28	[-74.67, 110.33]	65
Feminine	130.37*	48.78	[32.18, 225.58]	100
p(regress)				
Incorrect	-0.58	0.58	[-1.85, 0.42]	14
Mismatch vs. Match	-0.36	0.55	[-1.46, 0.71]	25
Single vs. Double	-0.34	0.47	[-1.33, 0.55]	23
Feminine	-0.24	0.44	[-1.12, 0.60]	30
p(skip)				
Incorrect	-0.02	0.25	[-0.51, 0.48]	47
Mismatch vs. Match	-0.53*	0.26	[-1.05, -0.02]	2
Single vs. Double	0.28	0.23	[-0.17, 0.73]	90
Feminine	-0.52*	0.22	[-0.94, -0.10]	1

5.3.3.3. Incorrect ORCs versus Correct SRCs

We found significant effects for Clause Type at the relative clause verb for go-past and total fixation duration, such that misinterpreted ORCs were read slower than correct SRCs in this region. However, we found no significant effects for any Grammatical Condition on any fixation metric.

We also found a significant effect for Clause Type for first pass skips the relative clause NP, such that misinterpreted ORCs were less likely to be skipped on first pass than SRCs in this region. We did not find any other significant effects by Grammatical Condition.

5.4. Discussion

This study set out to investigate whether the number of grammatical cues given in favor of a low-frequency interpretation can overcome expectations in a good-enough or noisy-channel framework. We conducted a second eye-tracking experiment with Arabic relative clauses, and added a grammatical manipulation to the stimuli to create mismatches between the matrix clause NP and relative clause NP in terms of grammatical number and gender. Experiment 3 utilized items where the matrix and relative clause NPs matched in both number and gender, meaning that the relative clause could not be disambiguated by grammatical marking alone; rather, the only way to distinguish between an SRC and an ORC was by reading and processing the resumptive object pronoun clitic on the ORC verb. By adding grammatical mismatches between these two NPs, we provided additional cues in favor of an ORC interpretation and against a noisy SRC interpretation, and tested whether these additional cues had any impact on both processing behaviors and accuracy rates in ORCs.

We started by analyzing the comprehension question answers to see whether the different grammatical manipulations had any effect on the likelihood of accepting a good-enough or noisy interpretation (i.e., selecting the incorrect comprehension question answer). We hypothesized that adding additional cues in favor of a low-frequency interpretation would decrease the number of noisy interpretations, as the reader would have to make an increased number of "edits" to arrive at their

preferred interpretation. We also hypothesized that this decrease in noisy interpretations would be proportional to the number of cues added: the Single Mismatch condition would have fewer noisy interpretations than the Match condition, and the Double Mismatch condition would have even fewer noisy interpretations than the Single Mismatch condition.

We found that there were significant differences in accuracy rates by Clause Type, such that ORCs were more likely to receive incorrect comprehension question answers than SRCs. However, there were no significant interactions between Clause Type and Grammatical Condition, meaning that these additional grammatical cues did not have a significant effect on the likelihood of a noisy interpretation of an ORC. There were also no significant differences by any grammatical condition across clause types. Thus, these additional grammatical cues were not strong enough to overcome the overwhelming expectation of an SRC over an ORC. These results align with similar work done in Hebrew: readers were more likely to accept a noisy SRC interpretation than a rare ORC interpretation, despite grammatical cues for both number and gender at the relative clause verb pointing toward an ORC interpretation (Keshev & Meltzer-Asscher, 2021). However, Keshev & Meltzer-Asscher tested two types of ORCs – one relatively rare and one more common – and found that this effect did not hold for the more common ORC structure. When comparing a noisy SRC to the more common ORC structure, readers were more likely to choose the common ORC over the noisy SRC, and thus, readers appeared to accept more noise in the input when faced with a very low probability structure compared to a much higher probability structure. Future research could investigate how the probability of a specific type of structure may modulate the use of grammatical cues in Arabic.

Since we found no significant interaction between Clause Type and Grammatical Condition on accuracy rates, we decided to conduct an omnibus test to investigate the possibility of significant interactions in our reading data. We started by investigating significant interactions between Clause Type and Grammatical Condition during veridical processing by comparing correctly interpreted ORCs to correctly interpreted SRCs, specifically within the relative clause verb and NP regions.

Model estimates revealed that there were no significant interactions between Clause Type and Grammatical Condition in these regions during veridical processing, meaning that the various grammatical manipulations did not have any significant effect on reading behaviors in ORCs in these critical regions. These results align with our findings from the comprehension question analysis, which demonstrated that Grammatical Condition also did not have a significant effect on the likelihood of a good-enough or noisy interpretation. They further show that these additional grammatical cues also do not trigger any sort of different reading strategy when processing ORCs. This is particularly interesting considering the temporary ambiguity of an ORC in the Match condition at the relative clause verb (see Section 1.4.2). In the Match condition, the relative clause verb agrees grammatically with both the matrix subject and relative clause NPs, meaning that an ORC with a relative clause verb with a resumptive object pronoun could temporarily be interpreted as an SRC with a resumptive pronoun and null object until the reader encounters the relative clause NP, thus eliminating that possible interpretation. In the Mismatch conditions, this interpretation is not possible: the relative clause verb does not agree grammatically with the matrix subject, and thus cannot be considered to be an SRC even temporarily. Given that this is a possible interpretation in the Match condition but not in either Mismatch condition, we might expect to see a significant interaction between Clause Type and Grammatical Condition at the relative clause verb. However, it is also possible that we do not have the statistical power to detect such an effect.

We also plotted average reading times in these regions by Clause Type and Grammatical Condition to examine whether we saw a numeric trend of an interaction, despite finding no statistical evidence for a significant interaction. ORC trials showed a numeric trend toward an interaction in early reading time measures at the relative clause verb, such that average first pass and go-past times were fastest for the Match condition and slowest for the Double Mismatch condition. However, later stage measures such as total fixation duration showed the longest average reading times for the Single Mismatch condition over the Double Mismatch condition. Reading times for ORCs at the relative clause NP showed a similar trend for average first pass, go-past, and total fixation duration times, with the Single Mismatch condition

having the longest average times. We cautiously interpret these results as suggesting that grammatical number may be a stronger cue than grammatical gender: the change from the Mismatch to Single Mismatch condition was a change in grammatical number, while the change from the Single Mismatch to Double Mismatch condition was a change in grammatical gender. Overall, our statistical analyses do not show any evidence of a significant interaction between Clause Type and Grammatical Condition, and we also find no evidence of this relationship numerically.

We then wanted to investigate the possibility of significant interactions within our misinterpreted ORC trials, as our first omnibus analysis only considered trials with correct comprehension question answers. We compared misinterpreted ORC trials to correctly interpreted ORC trials to test for a significant interaction between Correctness and Grammatical Condition, then compared misinterpreted ORCs to correctly interpreted SRC trials to again test for a significant interaction between Clause Type and Grammatical Condition. We fit two sets of models: one that included two-way interactions between Correctness or Clause Type and Grammatical Condition, and one that only considered the additive effects of these variables.

WAIC comparisons for models comparing misinterpreted ORCs to correct ORCs showed that the additive model without any interactions was the best fit for our data. Following the findings from our first omnibus analysis, there were no significant interactions between Correctness and Grammatical Condition, meaning that there were no significant behavioral differences in ORCs across grammatical conditions. This once again aligns with the findings from our comprehension question analysis, but additionally shows that there were no differences in reading behaviors when arriving at the correct interpretation of an ORC versus the incorrect interpretation of an ORC, regardless of grammatical condition.

WAIC comparisons for models comparing misinterpreted ORCs to correct SRCs showed more of a tie: for first pass and go-past times at the relative clause verb, the interaction model was the best fit, while for first pass regressions at the relative clause verb and total fixation durations at the relative clause NP, the additive model was the best fit. We decided to explore the significant interactions for first pass and

go-past times at the relative clause verb by plotting averages for these times by Clause Type and Grammatical Condition. These averages showed distinct patterns by Clause Type: SRC trials showed a slight increase in reading times across grammatical conditions, with items in the Match condition being read fastest and items in the Double Mismatch condition being read slowest, while ORC trials showed that items in the Single Mismatch condition were read slowest, above the Double Mismatch condition. These results provide tentative support for our previous speculation that number may be a stronger cue than gender when it comes to processing times. However, we still found no significant interactions between Correctness and Grammatical Condition when comparing misinterpreted ORCs to correct ORCs, so we did not explore this interaction further.

Overall, our omnibus analyses demonstrated that Grammatical Condition did not have a significant effect on processing behaviors in ORCs compared to SRCs, and also did not have a significant effect on reading behaviors that led to a correct interpretation of an ORC compared to an incorrect interpretation. Adding additional cues in favor of an ORC interpretation was not sufficient to alter reading strategies for ORCs, nor did it affect reading strategies for correctly interpreted ORCs compared to misinterpreted ORCs. Combined with the results from our comprehension question analysis, we take these results to show that grammatical cues such as number and gender agreement that show support for a lower-frequency interpretation are not strong enough to overcome expectations for a higher-frequency interpretation, and do not affect reading strategies when processing these structures.

Finally, we returned to our main analyses to investigate the individual effects of Clause Type and Grammatical Condition in our data. We tested whether we saw differences by Clause Type that matched our results from Experiment 3, and whether the different grammatical conditions affected processing behaviors overall, across clause types. We coded our grammatical mismatch conditions so that the results showed us how changing either grammatical number or grammatical gender independently affected processing behaviors: the Match condition items were compared to both Mismatch condition items to

determine differences by grammatical number, and the two Mismatch conditions (Single and Double) were compared to one another to determine differences by grammatical gender.

We first analyzed differences during veridical processing, between SRC and ORC trials that received correct comprehension question answers. We again found that ORCs were read slower than SRCs, in line with expectation-based theories of sentence processing and our previous findings. However, we found significant differences by Clause Type in both the relative clause verb and relative clause NP regions. ORCs had longer first pass, go-past and total fixation durations at the relative clause verb, and had longer go-past times and a higher likelihood of a first pass regression at the relative clause NP. These findings partially align to what we found in Experiment 3, that participants are paying a higher processing cost for ORCs specifically at the relative clause NP. However, we find that participants in Experiment 4 appear to be paying these costs at both the relative clause verb and relative clause NP.

Crucially, this effect is not dependent on Grammatical Condition. In Experiment 3 (where all items were in the Match condition), we hypothesized that we did not see an effect at the relative clause verb as participants could still be entertaining an SRC interpretation with an RP and a null object, until they reach the relative clause NP and that analysis is invalidated. This interpretation is not possible in the Mismatch conditions, as the relative clause verb no longer agrees grammatically with the matrix clause subject. So, finding this main effect for Clause Type could be due to one of two things. First, readers from this participant population could have a different reading strategy than those from Experiment 3. Participants from Experiment 4 were largely heritage speakers and made up a very heterogeneous sample from different dialects and different amounts of time living in Arabic-speaking countries, whereas participants from Experiment 3 were nearly all Emirati natives who were born and raised in the same country. Thus, the heritage speakers from Experiment 4 could be utilizing different reading strategies than the immersed speakers from Experiment 3 (e.g., AlQahtani & Sabourin, 2015). On the other hand, this main effect could be due simply to the fact that two-thirds of our items were in a Mismatch condition compared to a Match condition, so the main effect for Clause Type is relying heavily on behaviors in the Mismatch

conditions. In other words, it could be that there is a true interaction effect between Clause Type and Grammatical Condition, but we lack the power to detect it, and are therefore seeing an erroneous main effect of Clause Type that in reality only exists in the Mismatch conditions.

We additionally found significant differences by Grammatical Condition at the relative clause verb and NP. Items that were grammatically plural had longer total fixation duration times than items that were grammatically singular at the relative clause verb. Further, adding a secondary cue of gender from the Single to the Double Mismatch condition caused even longer first pass times at the relative clause verb. However, we did not see any significant differences by Grammatical Condition at the relative clause NP. This shows that overall, Match items demonstrate a sort of facilitative ambiguity at the relative clause verb (Levy, 2008a), where the verb could agree with the matrix clause NP or the relative clause NP, while Mismatch items that create more and more differences between these NPs slow down processing.

This difference between the Match and Mismatch conditions may arise from the fact that participants must pay closer attention to the verb to arrive at the correct interpretation of the sentence, so this slowdown across conditions could be a learned strategy. On the other hand, these results could suggest that verbs that are inflected to agree with grammatically plural subjects are more difficult to process than those that are grammatically singular, and creating a grammatical gender contrast between the matrix clause NP and relative clause NP additionally slows down processing at the verb. This second finding is interesting because this effect is not modulated by Clause Type: this effect is not specifically due to reading a relative clause verb that does not agree with the matrix clause subject, as would be the case in an ORC. Further, at this point in the sentence the reader has not yet encountered the relative clause NP and cannot know that there is a grammatical gender mismatch between the nouns. It is possible that they may receive some information from the relative clause NP parafoveally, but Arabic marks gender at the end of a word, so it is unlikely that they could register that gender marking parafoveally. Overall, we find that creating additional mismatches between the matrix clause and relative clause NPs slows down

processing at the relative clause verb; however, questions remain as to whether this difficulty is due to the inherent grammatical properties of the verb, or participant- and task-specific reading strategies.

We then analyzed the differences in processing behaviors in misinterpreted ORC trials compared to both correct ORC and correct SRC trials to investigate how various grammatical conditions affect processing behaviors during good-enough or noisy-channel processing. We found one significant difference between misinterpreted and correct ORCs at the relative clause verb for first pass skip rates, but found no significant differences in reading times or regression rates in this region. This aligns with our findings from Experiment 3 and suggests that readers were not misreading the verb when misinterpreting ORCs, supporting our original findings of good-enough or noisy-channel processing. Rather, differences between misinterpreted and correct ORCs manifested at the relative clause NP, where correct ORCs had a higher probability of a first pass regression. Notably, this is the same region where correct ORCs incur the most processing difficulty relative to correct SRCs and once again aligns with our findings from Experiment 3.

We also found significant differences by Grammatical Condition at the relative clause verb, similar to those found in veridical processing. Items that were grammatically plural had longer go-past and total fixation durations than items that were grammatically singular at the relative clause verb. However, we did not find the same difference by grammatical gender as in veridical processing. These results show that, within ORCs specifically, mismatching items are read slower across the board.

On the other hand, misinterpreted ORCs and correct SRCs had significant differences at the relative clause verb based on Clause Type: misinterpreted ORCs had longer go-past times and total fixation durations than correct SRCs. This aligns with the general differences observed between SRCs and ORCs during veridical processing for this experiment, though it does not directly align with the specific behaviors observed in Experiment 3. Further, misinterpreted ORCs and correct SRCs had no significant differences in reading times or first pass regressions at the relative clause NP. So, misinterpreted ORCs behave similarly to correct ORCs at the relative clause verb, but similarly to correct SRCs at the relative

clause NP. Overall, this pattern of results matches what we observed in Experiment 3 and again supports a good-enough or noisy-channel processing theory, rather than simple misreading by skipping the resumptive pronoun clitic. Finally, there were no significant differences in either region by Grammatical Condition, demonstrating that the addition of mismatching grammatical cues had no effect on distinguishing processing behavior between misinterpreted ORCs or correct SRCs at either the relative clause verb or relative clause NP.

One limitation of this research that has been addressed throughout this chapter is the number of datapoints from which we are drawing these conclusions. Native Arabic speakers who are proficient in MSA are difficult to come by in Davis, California, and so our data are sparser than we would have hoped. We hope to continue data collection in the future with more participants and more experimental items in order to further explore the questions and findings addressed here.

Overall, our results confirm the broad findings from our previous experiment: that Arabic readers pay an additional processing cost for ORCs at the relative clause NP, either instead of or in addition to any cost paid at the relative clause verb, in spite of the relative clause verb providing a very strong cue in favor of an ORC interpretation. Crucially, this processing cost at the relative clause NP is also the determining factor between correctly interpreting an ORC or misinterpreting an ORC, while processing behavior at the relative clause verb was *not* determinative of getting the correct interpretation. So, while the participants from Experiment 4 appear to pay more attention at the relative clause verb than the participants in Experiment 3, it is still necessary for the reader to pay some sort of processing cost at the relative clause NP in order to arrive at the correct interpretation, regardless of how much time they spend at the relative clause verb. Finally, adding additional grammatical cues such as a mismatch in grammatical number and gender does affect processing difficulty, but it does not affect the processing behavior crucial for correctly interpreting an ORC, nor does it affect the likelihood of accepting a good-enough or noisy interpretation. These results thus show that increasing the number of grammatical cues in

favor of a low-frequency interpretation is not sufficient to overcome expectations in a good-enough/noisy-channel framework.

Chapter 6

General Discussion

These studies set out to explore patterns of language processing and comprehension in relative clauses in Modern Standard Arabic. We tested whether observed processing difficulties in Arabic subject- and object-extracted relative clauses (SRCs and ORCs) were best explained by memory- or expectation-based theories of sentence processing and investigated the underlying causes of extensive misinterpretations of ORCs. Our first experiment, a self-paced reading task, showed support for expectation-based theories of sentence processing, indicating that ORCs are read slower than SRCs. However, this slowdown was significant not in the probabilistic disambiguating region (the relative clause verb), which was predicted based on a corpus analysis. Rather, it was significant in the following region, the relative clause NP. This result was tentatively attributed to spillover effects from the self-paced reading modality, but also potentially demonstrated that readers wait to update their expectations of a relative clause type until reading the relative clause NP. Results from our comprehension question analysis also revealed that participants were significantly more likely to misinterpret ORCs than SRCs, suggesting that readers were either misreading ORCs, or reading them correctly and instead accepting noisy SRC interpretations. We investigated this phenomenon first through a recall task, and then through an eye-tracking study.

Our eye-tracking study replicated the findings of our self-paced reading task, again showing support for expectation-based theories of sentence processing. The results also again revealed that the locus of

processing difficulty in ORCs is at the relative clause NP, and not at the relative clause verb where readers should be able to probabilistically disambiguate between an SRC and an ORC. We then compared processing behavior in trials where participants misinterpreted ORCs to trials where they correctly interpreted ORCs, and found that paying an extra integration processing cost at the relative clause NP was crucial to arriving at the correct interpretation of an ORC. These findings lend support to the notion of good-enough and noisy-channel processing, suggesting that readers may accept a noisy but more frequent SRC interpretation of ORCs to avoid the integration cost at the relative clause NP.

These results challenge some aspects of existing expectation-based theories of sentence processing, as they do not accurately predict the incremental processing patterns found in our results. According to expectation-based theories grounded in surprisal theory, readers should probabilistically adjust their expectations for upcoming input based on distributional statistics in the language. Specifically, when readers encounter a verb within an ORC that has a resumptive object pronoun clitic attached, they are statistically more likely to expect a noun phrase to follow, confirming an ORC interpretation. However, our results indicate that readers prioritize the global expectation of a SRC over an ORC, rather than making statistical inferences based on the specific structure of the clause. This suggests that Arabic speakers' language processing strategies are optimized to the language as a whole but are not as granular for the possibilities within a given structure. Existing expectation-based theories do not account for this global expectation taking precedence over a localized structural expectation.

One study on noisy-channel processing in relative clauses in Hebrew investigated how the frequency of the specific structure of an ORC might affect a reader's willingness to accept a noisy interpretation (Keshev & Meltzer-Asscher, 2021). The researchers used two forms of an ORC, one with a common structure and one with a rare structure, and tested participants' willingness to accept a noisy SRC interpretation. They found that participants chose the noisy SRC over the rare ORC, but preferred the common ORC over the noisy SRC. In this case, the expectation for a given structure appeared to outweigh the global expectation for an SRC. Our study did not include a manipulation for different

structures within one clause type, and the effect we discuss here is a *temporary* preference for a rare SRC until that structure is ruled out upon reading the relative clause NP, so we do not attempt to draw a direct comparison to these results. However, given these findings, a compelling line for future research could investigate the features that may contribute to prioritizing a global expectation over a structural expectation, and vice versa.

While an adjustment to expectation-based theories could explain our results, the potential effects of memory-based constraints remain a question. Memory-based theories state that readers should resolve long-distance dependencies as soon as possible; thus, for Arabic ORCs, resolving the matrix clause subject dependency at the relative clause verb would be advantageous from a cognitive standpoint. However, Arabic comprehenders do not follow this traditionally resource-rational strategy. One possible explanation could arise from joint theories of language processing, such as noisy- or lossy-context surprisal (Futrell et al., 2020; Futrell & Levy, 2017). Lossy-context surprisal posits that a comprehender's experience with distributional statistics in a particular language influences memory effects for that language. This effect has been demonstrated with structural forgetting phenomena such as grammaticality illusion. For example, English speakers often struggle with double-embedded relative clauses and may incorrectly judge ungrammatical sentences as grammatical due to forgetting embedded structures (Frazier, 1985; Gibson & Thomas, 1999). Surprisingly, this effect does not occur in German, regardless of it being the same exact structure (Vasishth et al., 2010). Authors theorized that since German has a higher frequency of verb-final structures, it makes double-embedded relative clauses with multiple verb-final clauses easier to process. In Arabic ORCs, memory-based theories grounded in DLT would suggest that holding an unresolved dependency for longer than necessary is disadvantageous and that dependencies should be resolved quickly. However, lossy-context surprisal shows that experience with a language's distributional statistics can potentially modulate proposed difficulty from memory constraints. Arabic's default word order is VSO, meaning that Arabic readers are accustomed to reading numerous post-verbal arguments. So, it is possible that experience with VSO word order helps to modulate the theoretical added

processing cost of waiting until after the relative clause verb to resolve the matrix clause subject dependency.

We then explored how grammatical and/or morphological cues affect a reader's willingness to accept a good-enough or noisy interpretation. In Experiment 3, the relative clause verb agreed grammatically with both the matrix and relative clause NP, meaning that the resumptive pronoun clitic provided the only signal of an ORC interpretation. Further, the resumptive pronoun clitics differed in length from one character to two characters. Our final experiment further investigated good-enough and noisy-channel processing in this area by testing the tradeoff of the strength of grammatical cues with the strength of expectations. We tested this question by conducting another eye tracking study where we manipulated the number and gender of the matrix and relative clause nouns to create a mismatch between the two. By creating a mismatch between the nouns, the relative clause verb only agreed grammatically with the relative clause noun and the resumptive pronoun only agreed grammatically with the matrix clause noun, providing additional signals for an ORC interpretation. We predicted that in the case of good-enough/noisy-channel processing, if increased grammatical cues mediate the acceptability threshold of noisy interpretations, the findings would show less noisy interpretations as the mismatch signals get stronger.

The results from this final experiment showed that an increased number of grammatical cues in favor of a veridical interpretation had no significant effect on the number of noisy interpretations of ORCs, nor on the reading strategies used while comprehending ORCs. Thus, readers were willing to accept ungrammatical but preferred SRC interpretations over a grammatical but less preferred ORC interpretation (Frazier, 1987; Keshev & Meltzer-Asscher, 2021). Our findings indicate that readers utilize sophisticated probabilistic knowledge about the distribution of structures in their language while processing text in real-time. Consequently, the frequency of different syntactic structures influences how readers interpret bottom-up cues, such as an agreement mismatch, leading to the adoption of noisy or

"good-enough" interpretations. Thus, grammatical cues are not strong enough to overcome structural expectations.

What remains to be explored in this area is what *does* affect statistical expectations within a good-enough or noisy-channel processing framework. One element that we did not explore as part of this research program is the effect of *semantic* expectations on *syntactic* expectations. Whereas our previous experiments compared different types of syntactic expectations, future research could broaden this scope to consider the interaction of plausibility and syntactic expectations: does strengthening semantic cues for a low-frequency interpretation mediate syntactic expectations of a high-frequency interpretation? For example, we could use similar SRC and ORC stimuli as the previous experiments, but modify the end of the sentence to provide more or less semantic evidence for a given interpretation. The stimuli would have two conditions: a baseline condition in which either an SRC or ORC interpretation is possible, and one in which only an SRC or ORC interpretation is highly plausible. In this high plausibility condition, the end of the sentence (i.e., the verb phrase for the matrix clause) would be semantically manipulated to provide stronger evidence for an SRC or ORC interpretation of the relative clause. For example, in the sentence, "The reporter who the senator killed was honored at a memorial service on Monday," the overall meaning of the sentence heavily suggests an ORC interpretation of the relative clause as opposed to an SRC interpretation. Sentences in the baseline plausibility condition would have vague matrix clause verbal phrases such that either an SRC or ORC interpretation is possible (e.g., "The reporter who the senator killed was well known among his peers"). If semantic cues are strong signals against good-enough/noisy interpretations, we would expect to see much fewer noisy interpretations in the semantically manipulated condition than in the baseline condition within clause type. On the other hand, if semantic cues are not strong enough signals against good-enough/noisy interpretations, we would see comparable amounts of noisy interpretations across plausibility conditions within clause type. Further investigating these tradeoffs will help to provide further insight into the probabilistic nature of human language processing.

References

Abu–Rabia, S. (2001). The role of vowels in reading Semitic scripts: Data from Arabic and Hebrew. *Reading and Writing*, *14*(1), 39–59. https://doi.org/10.1023/A:1008147606320

Abu–Rabia, S. (2002). Reading in a root–based–morphology language: The case of Arabic. *Journal of Research in Reading*, *25*(3), 299–309. https://doi.org/10.1111/1467-9817.00177

Albirini, A. (2019). Why Standard Arabic Is Not a Second Language for Native Speakers of Arabic. *Al-'Arabiyya*, *52*, 49–72.

AlQahtani, S., & Sabourin, L. (2015). Syntactic Processing of Subjects in Different Word Orders in Arabic: Do Arabic Heritage speakers differ from Native speakers when processing SVO/VSO order? *Proceedings of the Annual Meeting of the Canadian Linguistic Association*, 1–15.

Alresaini, S. (2012). *Acquisition of Modern Standard Arabic by Speakers of Different Arabic Colloquial Varieties* [Unpublished doctoral dissertation]. The University of York.

Alresaini, S. (2016). Acquisition of Modern Standard Arabic by Speakers of Different Arabic Colloquial Varieties: Resumption in Object Relative Clauses. *Arab World English Journal*, *7*(4), 202–224. https://doi.org/10.24093/awej/vol7no4.14

Al-Sharafi, Y. M., & Gubaily, M. A. (2023). Investigating the Null Object in Arabic Language. *(Abhath) مجلة أبحاث*, *10*(1), 726–748.

Aoun, J. E., Benmamoun, E., & Choueiri, L. (2010). *The Syntax of Arabic*. Cambridge University Press.

Barr, D. J., Roger, L., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bock, J. K., & Brewer, W. F. (1974). Reconstructive recall in sentences with alternative surface structures. *Journal of Experimental Psychology*, *103*(5), 837–843. https://doi.org/10.1037/h0037391

Boudelaa, S., & Marslen-Wilson, W. D. (2001). Morphological units in the Arabic mental lexicon. *Cognition*, *81*(1), 65–92. https://doi.org/10.1016/S0010-0277(01)00119-6

Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for Modern Standard Arabic. *Behavior Research Methods*, *42*(2), 481–487. https://doi.org/10.3758/BRM.42.2.481

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, 104047. https://doi.org/10.1016/j.jml.2019.104047

Bürkner, P. C. (2017). Advanced Bayesian multilevel modeling with the R package brms. *R Journal*, *10*(1), 395–411. https://doi.org/10.32614/rj-2018-017

Carreiras, M., Duñabeitia, J. A., Vergara, M., de la Cruz-Pavía, I., & Laka, I. (2010). Subject relative clauses are not universally easier to process: Evidence from Basque. *Cognition*, *115*(1), 79–92. https://doi.org/10.1016/j.cognition.2009.11.012

Central Intelligence Agency. (2018). World: People and Society. In *The World Factbook*. https://www.cia.gov/the-world-factbook/countries/world/#people-and-society

Chen, B., Ning, A., Bi, H., & Dunlap, S. (2008). Chinese subject-relative clauses are more difficult to process than the object-relative clauses. *Acta Psychologica*, *129*(1), 61–65. https://doi.org/10.1016/j.actpsy.2008.04.005

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic Roles Assigned along the Garden Path Linger. *Cognitive Psychology*, *42*(4), 368–407. https://doi.org/10.1006/cogp.2001.0752

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2017). Why reread? Evidence from garden-path and local coherence structures. *Quarterly Journal of Experimental Psychology*, *70*(7), 1380–1405. https://doi.org/10.1080/17470218.2016.1186200

Cutter, M. G., Filik, R., & Paterson, K. B. (2022). Do readers maintain word-level uncertainty during reading? A pre-registered replication study. *Journal of Memory and Language*, *125*(104336), 1–14. https://doi.org/10.1016/j.jml.2022.104336

Cutter, M. G., Paterson, K. B., & Filik, R. (2024). Eye-movements during reading and noisy-channel inference making. *Journal of Memory and Language*, *137*(104513), 1–16. https://doi.org/10.1016/j.jml.2024.104513

Demberg, V., & Keller, F. (2008). Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity. *Cognition*, *109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

Demberg, V., & Keller, F. (2009). A Computational Model of Prediction in Human Parsing: Unifying Locality and Surprisal Effects. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*(31), 1888–1893.

Demberg, V., Keller, F., & Koller, A. (2013). Incremental, Predictive Parsing with Psycholinguistically Motivated Tree-Adjoining Grammar. *Computational Linguistics*, *39*(4), 1025–1066. https://doi.org/10.1162/COLI_a_00160

Doornik, J. A., & Hansen, H. (2008). An Omnibus Test for Univariate and Multivariate Normality. *Oxford Bulletin of Economics and Statistics*, *70*(s1), 927–939. https://doi.org/10.1111/j.1468-0084.2008.00537.x

Ellis, N. C. (2002). Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/s0272263102002024

Farid, M., & Grainger, J. (1996). How initial fixation position influences visual word recognition: A comparison of French and Arabic. *Brain and Language*, *53*(3), 351–368. https://doi.org/10.1006/brln.1996.0053

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. https://doi.org/10.1111/1467-8721.00158

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*(3), 348–368. https://doi.org/10.1016/0749-596X(86)90006-9

Ferreira, F., & Lowder, M. W. (2016). Prediction, Information Structure, and Good-Enough Language Processing. In *Psychology of Learning and Motivation* (Vol. 65, pp. 217–247). Academic Press. https://doi.org/10.1016/bs.plm.2016.04.002

Ferreira, F., & Patson, N. D. (2007). The 'Good Enough' Approach to Language Comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Flores D'Arcais, G. B. (1974). Is there a memory for sentences? *Acta Psychologica*, *38*(1), 33–58. https://doi.org/10.1016/0001-6918(74)90028-6

Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, *45*(4), 1182–1190. https://doi.org/10.3758/s13428-012-0313-y

Frazier, L. (1985). Syntactic complexity. In *D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge University Press.

Frazier, L. (1987). Syntactic processing: Evidence from Dutch. *Natural Language & Linguistic Theory*, *5*(4), 519–559. https://doi.org/10.1007/BF00138988

Friederici, A. D., Steinhauer, K., Mecklinger, A., & Meyer, M. (1998). Working memory constraints on syntactic ambiguity resolution as revealed by electrical brain responses. *Biological Psychology*, *47*(3), 193–221. https://doi.org/10.1016/S0301-0511(97)00033-1

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, *44*(3), 1–54. https://doi.org/10.1111/cogs.12814

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 688–698. https://doi.org/10.18653/v1/E17-1065

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity. In *A. Marantz, W. A. O'Neil, & Y. Miyashita (Eds.), Image, Language, Brain* (pp. 95–126). MIT Press.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110

Gibson, E., & Thomas, J. (1999). Memory Limitations and Structural Forgetting: The Perception of Complex Ungrammatical Sentences as Grammatical. *Language and Cognitive Processes*, *14*(3), 225–248. https://doi.org/10.1080/016909699386293

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory Interference during Language Processing. *Journal of Experimental Psychology: Learning Memory and Cognition*, *27*(6), 1411–1423. https://doi.org/10.1037/0278-7393.27.6.1411

Grodner, D., & Gibson, E. (2005). Consequences of the Serial Nature of Linguistic Input for Sentenial Complexity. *Cognitive Science*, *29*(2), 261–290. https://doi.org/10.1207/s15516709cog0000_7

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), 1–9. https://doi.org/10.1073/pnas.2122602119

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1–8. https://doi.org/10.3115/1073336.1073357

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. OUP Oxford.

Hermena, E. (2016). *Aspects of word and sentence processing during reading Arabic: Evidence from eye movements* [Unpublished doctoral dissertation]. University of Southampton.

Hermena, E. W., Drieghe, D., Hellmuth, S., & Liversedge, S. P. (2015). Processing of Arabic diacritical marks: Phonological-syntactic disambiguation of homographic verbs and visual crowding effects. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 494–507. https://doi.org/10.1037/xhp0000032

Hermena, E. W., Liversedge, S. P., Bouamama, S., & Drieghe, D. (2019). Orthographic and root frequency effects in Arabic: Evidence from eye movements and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(5), 934–954. https://doi.org/10.1037/xlm0000626

Holes, C. (2004). *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press.

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*(1), 3–27. https://doi.org/10.1016/S0010-0277(03)00124-0

Huang, K. J., & Staub, A. (2021). Why do readers fail to notice word transpositions, omissions, and repetitions? A review of recent evidence and theory. *Language and Linguistics Compass*, *15*(7), 1–17. https://doi.org/10.1111/lnc3.12434

Huang, Y., & Ferreira, F. (2021). What causes lingering misinterpretations of garden-path sentences: Incorrect syntactic representations or fallible memory processes? *Journal of Memory and Language*, *121*(104228), 1–15.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, *40*(6), 431–439. https://doi.org/10.3758/BF03208203/METRICS

James, C. T., Thompson, J. G., & Baldwin, J. M. (1973). The reconstructive process in sentence memory. *Journal of Verbal Learning and Verbal Behavior*, *12*(1), 51–63. https://doi.org/10.1016/S0022-5371(73)80060-X

Jordan, T. R., Almabruk, A. A. A., Gadalla, E. A., McGowan, V. A., White, S. J., Abedipour, L., & Paterson, K. B. (2014). Reading direction and the central perceptual span: Evidence from Arabic and English. *Psychonomic Bulletin & Review*, *21*(2), 505–511. https://doi.org/10.3758/s13423-013-0510-4

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. https://doi.org/10.1037/0096-3445.111.2.228

Keshev, M., & Meltzer-Asscher, A. (2021). Noisy is better than rare: Comprehenders compromise subject-verb agreement to form more probable linguistic structures. *Cognitive Psychology*, *124*, 1–58. https://doi.org/10.1016/j.cogpsych.2020.101359

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*(5), 580–602. https://doi.org/10.1016/0749-596X(91)90027-H

Konieczny, L., & Döring, P. (2003). Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. *Proceedings of the ICCS/ASCS-2003 Joint International Conference on Cognitive Science*, 1–6.

Kwon, N., Lee, Y., Gordon, P. C., Kluender, R., & Polinsky, M. (2010). Cognitive and Linguistic Factors Affecting Subject/Object Asymmetry: An Eye-Tracking Study of Prenominal Relative Clauses in Korean. *Language*, *86*(3), 546–582.

Lau, E., & Tanaka, N. (2021). The subject advantage in relative clauses: A review. *Glossa: A Journal of General Linguistics*, *6*(1), 1–34. https://doi.org/10.5334/gjgl.1343

Leung, T., Ntelitheos, D., & Al Kaabi, M. (2020). *Emirati Arabic: A Comprehensive Grammar*. Routledge. https://doi.org/10.4324/9780429273162

Levy, R. (2008a). A noisy-channel model of rational human sentence comprehension under uncertain input. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 234–243.

Levy, R. (2008b). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: Formal techniques and empirical results. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1055–1065.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090. https://doi.org/10.1073/pnas.0907664106

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition*, *23*(5), 938–944. https://doi.org/10.1017/S1366728918001153

Maamouri, M., Bies, A., Kulick, S., Krouna, S., Gaddeche, F., & Zaghouani, W. (2010). *Arabic Treebank: Part 3 v 3.2 LDC2010T08* [Dataset]. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2010T08

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical Nature of Syntactic Ambiguity Resolution. *Psychological Review*, *101*(4), 676–703. https://doi.org/10.1037//0033-295x.101.4.676

Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. *Journal of Memory and Language*, *47*(1), 50–68. https://doi.org/10.1006/jmla.2001.2837

Meltzer-Asscher, A. (2021). Resumptive Pronouns in Language Comprehension and Production. *Annual Review of Linguistics*, *7*, 177–194. https://doi.org/10.1146/annurev-linguistics-031320-012726

Parkinson, D. B. (1981). VSO to SVO in Modern Standard Arabic: A Study in Diglossia Syntax. *Al-'Arabiyya*, *14*(1), 24–37.

Paterson, K. B., Almabruk, A. A. A., McGowan, V. A., White, S. J., & Jordan, T. R. (2015). Effects of word length on eye movement control: The evidence from Arabic. *Psychonomic Bulletin and Review*, *22*(5), 1443–1450. https://doi.org/10.3758/s13423-015-0809-4

Pollatsek, A., Bolozky, S., Well, A. D., & Rayner, K. (1981). Asymmetries in the perceptual span for Israeli readers. *Brain and Language*, *14*(1), 174–180. https://doi.org/10.1016/0093-934X(81)90073-0

Poppels, T., & Levy, R. P. (2016). Structure-sensitive Noise Inference: Comprehenders Expect Exchange Errors. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, *124*(3).

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348–379. https://doi.org/10.1016/j.jml.2007.03.002

Roman, G., & Pavard, B. (1987). A comparative study: How we read in Arabic and French. *Eye Movements from Physiology to Cognition*, 431–440. https://doi.org/10.1016/B978-0-444-70113-8.50064-3

Ryding, K. C. (2005). *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press.

Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders Model the Nature of Noise in the Environment. *Cognition*, *181*, 141–150. https://doi.org/10.1016/j.cognition.2018.08.018

Schriefers, H., Friederici, A. D., & Kuhn, K. (1995). The Processing of Locally Ambiguous Relative Clauses in German. *Journal of Memory and Language*, *34*, 499–520.

Sinnemäki, K. (2010). Word order in zero-marking languages. *Studies in Language*, *34*(4), 869–912. https://doi.org/10.1075/sl.34.4.04sin

Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Syntactic Ambiguity Resolution in Discourse: Modeling the Effects of Referential Context and Lexical Frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1521–1543.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86. https://doi.org/10.1016/j.cognition.2010.04.002

Thompson, E., & Werfelli, S. (2012). The Position of the Subject in Spoken Saudi Arabic: A Processing Perspective. *Coyote Papers: Working Papers in Linguistics at the University of Arizona*, *19*, 1–17.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*(1), 69–90. https://doi.org/10.1006/jmla.2001.2836

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language*, *53*(2), 204–224.

Ueno, M., & Garnsey, S. M. (2008). An ERP study of the processing of subject and object relative clauses in Japanese. *Language and Cognitive Processes*, *23*(5), 646–688. https://doi.org/10.1080/01690960701653501

Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, *25*(4), 533–567. https://doi.org/10.1080/01690960903310587

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, *11*(12), 3571–3594.

Appendices

Appendix A

Experimental stimuli:

[1] Items used in Experiment 1: Self-Paced Reading

[2] Items used in Experiment 2: Recall Task

[3] Items used in Experiment 3: Eye Tracking

[4] Items used in Experiment 4: Eye Tracking Part 2

* Items that were minorly changed for Experiment 4 to make sense in the number and gender mismatch conditions

Arabic	**English translation**
البنت التي (أيقظت/أيقظتها) الوالدة أزعجتها بشأن الرحلة إلى الشاطئ.	The girl who (woke the mother/the mother woke) bothered her about the trip to the beach. [1,2,3,4]
سائق الحافلة الذي (تبع/تبعه) الطفل تساءل عن موقع الفندق.	The bus driver who (followed the kid/the kid followed) wondered about the location of the hotel. [1,2,3,4]
القاضية التي (خاطبت/خاطبتها) الشاهدة لاحظت محامي الدفاع.	The judge who (addressed the witness/the witness addressed) noticed the defense attorneys. [1,2,3,4]
المدير الذي (زار/زاره) الرئيس تذكر بعض الحقائق غير المريحة.	The manager who (visited the boss/the boss visited) remembered some inconvenient facts. [1,2,3,4]
الجارة التي (لاحظت/لاحظتها) السمسارة اشترت المنزل القديم.	The neighbor who (observed the realtor/the realtor observed) purchased the old house. [1,2,3,4]
الطيار الذي (أخّر/أخّره) الطاقم الأرضي بقي على المدرج لفترة طويلة.	The pilot who (delayed the ground crew/the ground crew delayed) remained on the runway for a long time. [1,2,3]
المتحدثة التي (استضافت/استضافتها) الاقتصادية توقعت سنة جيدة لهذه الصناعة.	The speaker who (entertained the economist/the economist entertained) predicted a good year for the industry. [1,2,3,4]
الجندي الذي (أعجب/أعجبه) المدرب هزم أعظم منافسيه.	The soldier who (admired the coach/the coach admired) defeated his greatest rival. [1,2,3,4,*]
الزائر الذي (قدّم/قدّمه) الطالب مشى عبر الحرم الجامعي.	The visitor who (introduced the student/the student introduced) walked across the campus. [1,2,3,4]
المصرفي الذي (أغضب/أغضبه) المحامي لعب التنس كل يوم سبت.	The banker who (irritated the lawyer/the lawyer irritated) played tennis every Saturday. [1,2,3,4]

الطبيبة التي (تجاهلت/تجاهلتها) الممرضة قادت سيارة حمراء.	The doctor who (ignored the nurse/the nurse ignored) drove a red car. [1,2,3,4,*]
السجين الذي (هاجم/هاجمه) الحارس أثار الشغب.	The prisoner who (attacked the guard/the guard attacked) provoked the riot. [1,2,3,4]
المتجول الذي (تجاوز/تجاوزه) الصياد ضاع وكان يجب إنقاذه.	The hiker who (passed the fisherman/the fisherman passed) got lost and had to be rescued. [1,2,3,4]
المستأجرة التي (احتقرت/احتقرتها) المالكة اتصلت بالصحيفة لتقديم شكوى.	The tenant who (despised the landlord/the landlord despised) phoned the newspaper to complain. [1,2,3,4]
الأستاذة التي (انتقدت/انتقدتها) الطالبة خجلت وابتعدت.	The professor who (criticized the student/the student criticized) blushed and turned away. [1,2,3]
العميلة التي (واجهت/واجهتها) عالمة النفس هاجمتها في الليل.	The client who (confronted the psychologist/the psychologist confronted) attacked her in the night. [1,2,3,4]
المؤرخة التي (انتقدت/انتقدتها) مديرة المتحف غادرت المتحف فجأة.	The historian who (criticized the curator/the curator criticized) left the museum abruptly. [1,2,3,4]
الممرضة التي (عيّنتْ/عيّنتها) المساعدة درست في جامعة كامبريدج.	The nurse who (hired the assistant/the assistant hired) studied at Cambridge University. [1,2,3,4,*]
المهندس المعماري الذي (أحب/أحبه) رجل الإطفاء سيطر على المحادثة بينما كانت المباراة على شاشة التلفزيون.	The architect who (liked the fireman/the fireman liked) dominated the conversation while the game was on television. [1,2,3,4]
المصرفية التي (امتدحت/امتدحتها) المحللة تسلّقت الجبل قبل أن يتساقط الثلج.	The banker who (praised the analyst/the analyst praised) climbed the mountain before it snowed. [1,2,3,4]
الشاعرة التي (صادقت/صادقتها) الكاتبة كتبت سيرة ذاتية بعد أن أصبحت صداقتهما معروفة جيداً.	The poet who (befriended the author/the author befriended) wrote an autobiography after their friendship became well known. [1,2,3,4,*]
الخياط الذي (وصف/وصفه) العميل عمِل في متجر صغير بالقرب من محطة الحافلات.	The tailor who (described the customer/the customer described) worked in a small shop near the bus station. [1,2,3,4]
المدرب الذي (انتقد/انتقده) الحكم تحدث علنا عن الحادثة بعد المباراة.	The coach who (criticized the referee/the referee criticized) talked publicly about the incident after the game. [1,2,3]
المعيدة التي (كرهت/كرهتها) المعلمة قامت بتقليص قراءة الأسبوع.	The teaching assistant who (disliked the teacher/the teacher disliked) skimmed the reading for the week. [1,2,3,4]
الراقص الذي (أحب/أحبه) الجمهور تجاهل بعض المبادئ الأساسية.	The dancer who (loved the audience/the audience loved) ignored some basic principles. [1,2,3,4]
الموظف الذي (لاحظ/لاحظه) رجل الإطفاء سارع عبر الحقول المفتوحة.	The employee who (noticed the fireman/the fireman noticed) hurried across the open field. [1,2,3,4]
الفلاح الذي (قابل/قابله) الزبون رفع الدجاج من حظيرته.	The farmer who (contacted the customer/the customer contacted) lifted the chickens from their coop. [1,2,3]
عالم الرياضيات الذي (زار/زاره) رئيس مجلس الإدارة ابتكر حلاً لمشكلة معروفة.	The mathematician who (visited the chairman/the chairman visited) created a solution to the well-known problem. [1,2,3,4]

الممثلة المشهورة التي (زارت/زارتها) المُنظِمَة اقترحت جائزة سنوية.	The celebrity who (visited the organizer/the organizer visited) proposed an annual prize. [1,2,3,4]
الفتاة التي (شاهدت/شاهدتها) الأم غيرت جزءا مهماً من القصة.	The girl who (watched the mom/the mom watched) changed a critical part of the story. [1,2,3]
الفلاح الذي (استأجر/استأجره) المزارع قام بتكديس البذور في صفوف طويلة.	The farmer who (hired the rancher/the rancher hired) piled the seeds in long rows. [1,2,3,4]
الجندي الذي (ساعد/ساعده) المدني تسلق الصخرة الكبيرة التي سدّت الطريق.	The soldier who (helped the civilian/the civilian helped) climbed the big rock that blocked the path. [1,2,3,4]
المدرب الذي (ساعد/ساعده) الفارس فرك جلد الحصان.	The trainer who (helped the jockey/the jockey helped) rubbed the horse's skin. [1,2,3]
اللاعب الذي (ضرب/ضربه) حارس المرمى وقّع عقداً جديداً.	The player who (hit the goalkeeper/the goalkeeper hit) signed a new contract. [1,2,3,4,*]
الكاتبة التي (أغضبت/أغضبتها) المحررة كتبت مقالا احتجاجيا.	The writer who (angered the editor/the editor angered) wrote an article in protest. [1,2,3]
السباك الذي (ساعد/ساعده) الكهربائي تقاعد بعد عشرين عاما في العمل.	The plumber who (helped the electrician/the electrician helped) retired after twenty years on the job. [1,2,3]
الصياد الذي (رأى/رآه) الناشط هرب إلى الغابة.	The hunter who (saw the activist/the activist saw) ran off into the forest. [1,2,3,4]
الممثل الذي (زار/زاره) المخرج طالب بدور البطولة في الفيلم.	The actor who (visited the director/the director visited) demanded the starring role in the movie. [1,2,3,4,*]
القاضية التي (تجاهلت/تجاهلتها) الطبيبة شاهدت البرنامج عن تجار المخدرات الكولومبيين على الأخبار المسائية.	The judge who (ignored the doctor/the doctor ignored) watched the special about Colombian drug dealers on the nightly news. [1,2,3,4]
العمة التي (تسلي/تسليها) الفتاة قامت بصنع دمى ورقية من الصحيفة.	The aunt who (amused the girl/the girl amused) made paper dolls out of the newspaper. [1,2,3]
الصحفي الذي (هاجم/هاجمه) السيناتور اعترف بارتكاب خطأ.	The reporter who (attacked the senator/the senator attacked) admitted to making an error. [3,4]
الموسيقية التي (أهانت/أهانتها) مذيعة الأخبار غادرت المبنى بعد المقابلة.	The musician who (insulted the newscaster/the newscaster insulted) left the building after the interview. [3,4]
المتدرب الذي (أربك/أربكه) العالم عمل في مختبر شهير في جامعة هارفارد.	The intern who (confused the scientist/the scientist confused) worked at a famous lab at Harvard University. [3,4]
المحقق الذي (تبع/تبعه) الضابط حل قضايا مماثلة في الماضي.	The detective who (followed the officer/the officer followed) solved similar cases in the past. [3,4]
زميلة العمل التي (استقبلت/استقبلتها) السكرتارية أحضرت بعض الزهور إلى المكتب.	The coworker who (greeted the secretary/the secretary greeted) brought some flowers to the office. [3,4]
طبيبة الأطفال التي (أوصت/أوصتها) طبيبة الأسنان تركت رسالة عن الجرعة الموصى بها للمريض.	The pediatrician who (recommended the dentist/the dentist recommended) left a message about the recommended dosage for the patient. [3,4]

طبيب الأعصاب الذي (ساعد/ساعده) الممرض عمل في المستشفى المحلي على مدى السنوات العشر الماضية.	The neurologist who (helped the nurse/the nurse helped) worked at the local hospital for the last ten years. [3,4]
المخترع الذي (مدح/مدحه) الباحث حصل على براءات اختراع عدة خلال السنوات الخمس الماضية.	The inventor who (praised the researcher/the researcher praised) patented several inventions over the last five years. [3]
السفير الذي (زار/زاره) المترجم عاش في الجزائر لسنوات عديدة.	The ambassador who (visited the interpreter/the interpreter visited) lived in Algeria for many years. [3]
المديرة التي (شاهدت/شاهدتها) الموظفة كانت لديها مشكلة مع الشركة.	The manager who (watched the employee/the employee watched) had a problem with the company. [3,4]
عالمة الاجتماع التي (تحدّت/تحدّتها) عالمة الأنثروبولوجيا نشرت كتابا مشهورا حول نفس الموضوع.	The sociologist who (challenged the anthropologist/the anthropologist challenged) published a famous book on the same topic. [3,4]
المستشارة التي (انتقدت/انتقدتها) المرشدة ندمت على التعليق بعد العرض.	The counselor who (criticized the instructor/the instructor criticized) regretted the comment after the presentation. [3]
المحلل الذي (استشار/استشاره) المقاول كان متورطا في فضيحة منذ وقت ليس ببعيد.	The analyst who (consulted the manufacturer/the manufacturer consulted) was involved in a scandal not long ago. [3]
المهندس الذي (ألهم/ألهمه) عالم الرياضيات حصل مؤخرا على جائزة في مؤتمر.	The engineer who (inspired the mathematician/the mathematician inspired) received an award at a conference recently. [3,4]
الفنانة التي (أزعجت/أزعجتها) الكاتبة غادرت المتحف بمزاج سيء.	The artist who (bothered the writer/the writer bothered) left the museum in a bad mood. [3]
عازفة الجيتار التي (رافقت/رافقتها) المغنية بقيت على خشبة المسرح لتؤدي أغنية أخرى.	The guitarist who (accompanied the singer/the singer accompanied) stayed on stage to perform one more song. [3,4]
النجار الذي (ساعد/ساعده) عامل البناء كان لديه عشرين عاما من الخبرة.	The carpenter who (assisted the construction worker/the construction worker assisted) had twenty years of experience. [3,4]
المشرّع الذي (حذر/حذره) المحامي اتهم القاضي بأخذ رشاوى.	The legislator who (cautioned the lawyer/the lawyer cautioned) accused the judge of taking bribes. [3,4]
السكرتارية التي (صدقت/صدقتها) المديرة بلغت عن سوء المعاملة في المكتب.	The secretary who (believed the executive/the executive believed) reported mistreatment in the office. [3,4]
الوسيط العقاري الذي (اكتشف/اكتشفه) صاحب المنزل رتب اجتماعا لوضع اللمسات الأخيرة على الصفقة.	The realtor who (discovered the homeowner/the homeowner discovered) arranged a meeting to finalize the deal. [3,4]
الزبونة التي (كرهت/كرهتها) صاحبة المتجر سرقت الحلوى من المتجر.	The customer who (hated the store owner/the store owner hated) stole candy from the store. [3,4]
المحامي الذي (رأى/رآه) العميل غادر قاعة المحكمة ليذهب لمقابلة شخص ما.	The lawyer who (saw the client/the client saw) left the courthouse to go meet someone. [3,4]

المحقق الذي (استقبل/استقبله) المهندس عمل بدوام كامل في القضية.	The detective who (greeted the engineer/the engineer greeted) worked full-time on the case. [3,4]
بائعة الزهور التي (أحبت/أحبتها) البستانية رأت صورا لعملها في مجلة.	The florist who (adored the gardener/the gardener adored) saw photos of her work in a magazine. [3]
الشرطي الذي (واجه/واجهه) اللص استدعى دعما إضافيا من فريقه بسرعة.	The policeman who (encountered the thief/the thief encountered) quickly called for reinforcements from his team. [3,4]
الفنانة التي (امتدحت/امتدحتها) النحاتة عرضت لوحات في معرض الفنون المحلي.	The artist who (praised the sculptor/the sculptor praised) exhibited portraits at the local art gallery. [3]
المعلمة التي (فهمت/فهمتها) الطفلة علمت الرياضيات والعلوم للصف الرابع.	The teacher who (understood the child/the child understood) taught fourth grade math and science. [3,4]
المربية التي (أحبت/أحبتها) الأم لعبت مع الطفل طوال الصباح.	The nanny who (liked the mom/the mom liked) played with the child all morning. [3,4]
عازف الكمان الذي (مدح/مدحه) قائد الأوركسترا أهان المغني أثناء التمرين.	The violinist who (flattered the conductor/the conductor flattered) insulted the singer at the rehearsal. [3,4]
السباك الذي (لكم/لكمه) الرسام صرخ على النجار قبل المشاجرة.	The plumber who (punched the painter/the painter punched) yelled at the carpenter before the altercation. [3,4,*]
المحاسب الذي (نصح/نصحه) المهندس تحدث إلى السكرتير بعد الاجتماع.	The accountant who (advised the engineer/the engineer advised) spoke to the secretary after the meeting. [3,4]
الطالبة التي (اتهمت/اتهمتها) الأستاذة قابلت العميد قبل الفصل.	The student who (accused the professor/the professor accused) met with the dean before class. [3,4]
المصورة التي (تابعت/تابعتها) المصممة ابتكرت صوراً مميزة لحملات الموضة.	The photographer who (followed the designer/the designer followed) created iconic images for fashion campaigns. [3,4]
المجرم الذي (كره/كرهه) المحامي واجه القاضي باحتقار.	The criminal who (hated the lawyer/the lawyer hated) faced the judge with contempt. [3,4]
الخاطب الذي (أمتع/أمتعه) الملك أراد أن يرى الأميرة قبل الحفلة.	The suitor who (entertained the king/the king entertained) wanted to see the princess before the party. [3,4,*]
المتسابقة التي (احترمت/احترمتها) القاضية تحدثت مطولا إلى المصور.	The contestant who (respected the judge/the judge respected) spoke at length to the cameraman. [3]
الدبلوماسي الذي (أغضب/أغضبه) رئيس الوزراء غادر البلاد بعد مؤتمر القمة.	The diplomat who (angered the prime minister/the prime minister angered) left the country after the summit. [3,4]
السائحة التي (رافقت/رافقتها) المرشدة السياحية لوحت للراهبات خلال الزيارة.	The tourist who (accompanied the guide/the guide accompanied) waved at the nuns during the visit. [3,4]
الراكب الذي (عرف/عرفه) الطيار تحدث إلى طاقم الطائرة قبل الإقلاع.	The passenger who (knew the pilot/the pilot knew) talked to the cabin crew before takeoff. [3,4]
لاعب الجولف الذي (أرشد/أرشده) الزميل فاز بالبطولة الوطنية المرموقة.	The golfer who (mentored the teammate/the teammate mentored) won the coveted national championship. [3,4,*]

Sample comprehension question for Experiments 1 and 2:

Arabic	**English translation**
هل البنت أيقظت الوالدة؟ (a) نعم (b) لا	Did the child wake the mother? (a) Yes (b) No

Sample comprehension question for Experiments 3 and 4:

Arabic	**English translation**
أي من العبارات التالية صحيحة؟ (a) الصحفي هاجم السيناتور (b) السيناتور هاجم الصحفي	Which of the following statements is true? (a) The reporter attacked the senator (b) The senator attacked the reporter

Appendix B

Participant questionnaire for Self-Paced Reading (Chapter 2) and Recall Task (Chapter 3):

1. Age: (*open response*)
2. Gender: (*open response*)
3. Ethnicity: (*open response*)
4. Race: (*open response*)
5. Place of birth: (*open response*)
6. Where you grew up: (*open response*)
7. Native language: (*multi-select*)
    a. MSA
    b. Other (please specify): (*open response*)
8. Other languages spoken (please describe where learned & how long, degree of fluency/literacy, and how often used): (*open response*)
9. How proficient are you in MSA? (1 = poor, 10 = excellent) (*scale 1-10*)
10. How proficient are you in your dialect? (1 = poor, 10 = excellent) (*scale 1-10*)

Appendix C

Raw RTs from Self-Paced Reading (Chapter 2)

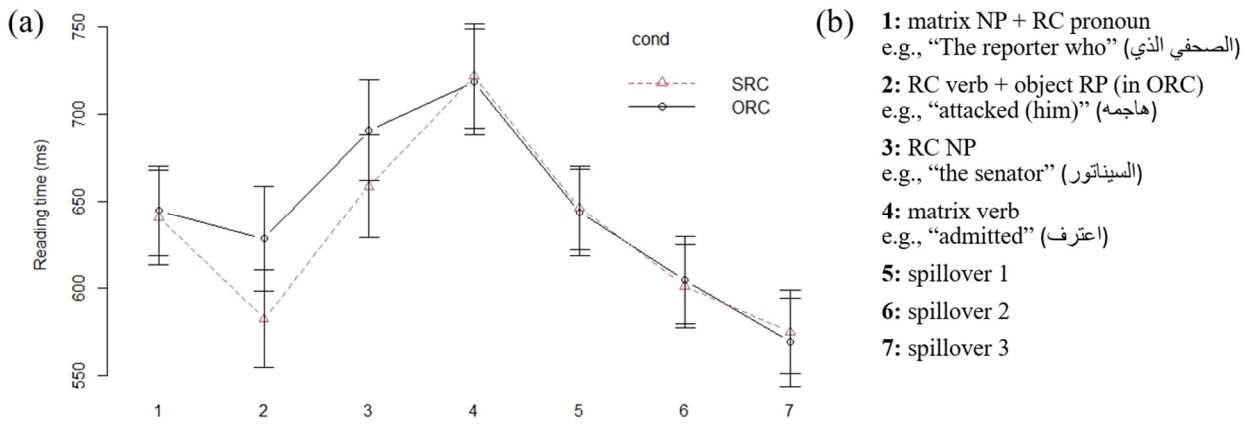

Figure C.1: (a) Average raw RTs for each region by clause type (after data preprocessing); (b) Regions of interest with Arabic examples and their English gloss.

Appendix D

Participant questionnaire for Eye Tracking experiment (Chapter 4):

1. Age: (*open response*)
2. Gender: (*single select*)
    a. Female
    b. Male
3. Country/Place of Birth: (*open response*)
4. Place where you grew up: (*open response*)
5. Duration of residence in the UAE (in years): (*open response*)
6. Do you have normal (or corrected-to-normal) vision? (*single select*)
    a. Yes
    b. No
7. What is your native language (mother tongue)? Please specify the particular dialect of the language (e.g. Emirati Arabic, Syrian Arabic, etc.): (*open response*)
8. For your native language listed above, please rate your current ability in terms of listening, speaking, reading, and writing (1 = very poor | 2 = poor | 3 = limited | 4 = average | 5 = good | 6 = very good | 7 = excellent) (*scale 1-7 for listening, speaking, reading, and writing each*)
9. Do you consider yourself to be proficient in Standard Arabic (SA)? (*single select*)
    a. Yes
    b. No
    c. Maybe
10. **If your answer was 'yes' OR 'maybe' for Standard Arabic**, please rate your current ability in terms of listening, speaking, reading, and writing (1 = very poor | 2 = poor | 3 = limited | 4 = average | 5 = good | 6 = very good | 7 = excellent) (*scale 1-7 for listening, speaking, reading, and writing each*)

11. How often do you use Standard Arabic? (*single select*)
    a. Always
    b. Often
    c. Sometimes
    d. Rarely
    e. Never
12. Excluding your native language and SA, what other languages do you speak? Please list them according to your proficiency in descending order. (*For each language listed, the participants then answered the following questions:*)
    a. Please rate your current ability in terms of listening, speaking, reading, and writing (1 = very poor | 2 = poor | 3 = limited | 4 = average | 5 = good | 6 = very good | 7 = excellent) (*scale 1-7 for listening, speaking, reading, and writing each*)
    b. How often do you use LANGUAGE? (*single select*)
        i. Always
        ii. Often
        iii. Sometimes
        iv. Rarely
        v. Never

Appendix E

Supplementary clitic analyses:

To investigate general fixation patterns on clitics in Arabic, 20 of the 80 filler items included in the experiment contained a clitic manipulation. One version of the filler would include an indefinite noun[8] (e.g., "The couple asked the realtor to visit a house in the neighborhood before making a decision."), and the other version would include the noun with a possessive pronoun in the form of a suffixed clitic (e.g., "The couple asked the realtor to visit their house in the neighborhood before making a decision."). Each possessive pronoun was either masculine singular (ـه; one character), feminine singular (ـها; two characters), or masculine plural (ـهم; two characters). Possessive pronouns for masculine and feminine singular were identical in surface form to the RP clitics included in the experimental stimuli. Table E.1 shows average fixation times and regression and skip rates by clitic type for ORC stimuli compared to our filler items.

Table E.1: Average reading times in ms and probabilities of regressions or skips by clitic type for ORC stimuli and filler items. No averages are included for the M.Pl clitic for ORC stimuli as none of them included this clitic.

Fixation Metrics	**M.Sg (ـه)**	**F.Sg (ـها)**	**M.Pl (ـهم)**
ORCs			
first fixation	47.59	63.42	-
first pass	48.10	68.00	-
go-past	65.16	105.16	-
total duration	167.42	215.04	-
p(regress)	0.25	0.24	-
p(skip)	0.82	0.77	-
Filler			
first fixation	43.57	78.58	107.81
first pass	44.83	78.58	115.52
go-past	74.11	125.54	148.87
total duration	91.65	144.51	179.21
p(regress)	0.17	0.20	0.15
p(skip)	0.85	0.71	0.66

[8] Indefinite nouns were specifically used instead of definite nouns because definite nouns in Arabic include a prefixed definite article clitic (e.g., البيت; al=bai:t; "the house") which is removed when a possessive pronoun is added (e.g., بيته; bai:t=uhu; "his house"). Indefinite nouns do not use indefinite articles (e.g., بيت; bai:t; "a house"), so adding a possessive pronoun to the indefinite noun would only require one edit, not two. Further, we wanted to avoid edits that would change the initial characters of the word to draw a more direct comparison to the fixation behaviors we were observing on the RC verb and RP clitic.

Appendix F

First-Half Second-Half analysis of Experiment 3 (Chapter 4):

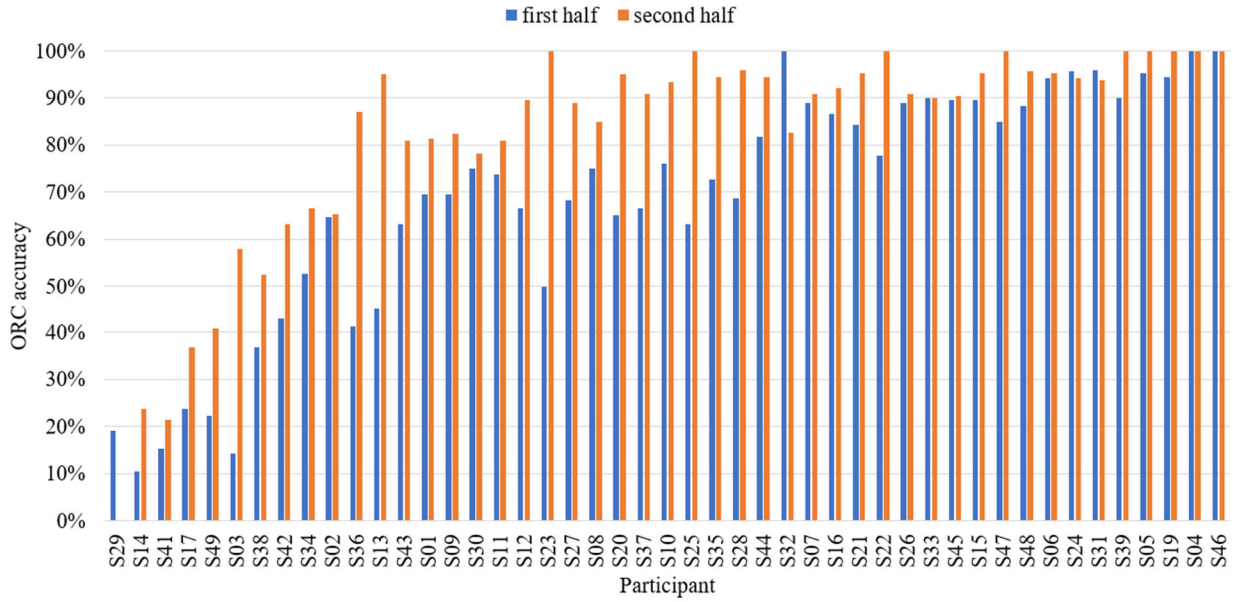

Figure F.1: Average ORC accuracy by participant for the first and second half of the experiment. Participants are ordered by overall average ORC accuracy, from lowest to highest.

Appendix G

Participant questionnaire for Experiment 4: Eye Tracking Part 2 (Chapter 5):

1. Age: (*open response*)
2. Gender: (*single select*)
    a. Male
    b. Female
    c. Non-binary, genderqueer, or gender non-conforming
    d. Prefer not to answer
3. Ethnicity: (*open response*)
4. Race: (*open response*)
5. Country/Place of Birth: (*open response*)
6. Place where you grew up: (*open response*)
7. Do you have normal (or corrected-to-normal) vision? (*single select*)
    a. Yes
    b. No
8. What is your native language (mother tongue)? Please specify the particular dialect of the language (e.g. Emirati Arabic, Syrian Arabic, etc.): (*open response*)
9. What other languages do you speak? Please list them according to your proficiency in descending order. (*Participants were instructed to list their native language first. For each language listed, the participants then answered the following questions:*)
    a. Please rate your current ability in terms of listening, speaking, reading, and writing (1 = very poor | 2 = poor | 3 = limited | 4 = average | 5 = good | 6 = very good | 7 = excellent) (*scale 1-7 for listening, speaking, reading, and writing each*)
    b. How often do you use LANGUAGE? (*single select*)
        i. Always

ii. Often

iii. Sometimes

iv. Rarely

v. Never

Appendix H

Full model estimates for Experiment 4: Eye tracking part 2 (Chapter 5)

Table H.1: Linear and logistic mixed-effects model estimates of the dependent variables on all fixation metrics at the matrix NP, RC pronoun, matrix verb, and spillover region 1 for the Cor ORC vs. Cor SRC analysis, including SE estimates and CrIs. The matrix NP region does not have an estimate for *p(regress)* as it is the first region in each sentence and a regressive saccade would be impossible. The % > 0 column shows the percent of the sampled posterior distribution that is over or under 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates that are bolded are significant.

Cor ORC vs. Cor SRC

	matrix NP				**RC pronoun**				**matrix verb**				**spillover 1**			
Fixation Metrics	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**
first fixation																
ORC	-4.87	6.40	[-17.52, 7.87]	22	-1.69	8.29	[-17.82, 14.72]	41	-2.17	9.48	[-20.82, 16.55]	41	**-14.75**	**9.10**	**[-32.54, 3.66]**	**5**
Mismatch vs. Match	6.82	7.21	[-7.25, 21.01]	83	-30.22	22.79	[-75.45, 13.91]	9	2.98	10.95	[-18.40, 24.41]	61	**17.34**	**10.85**	**[-3.41, 39.03]**	**95**
Double vs. Single	7.52	8.60	[-9.47, 24.45]	81	-9.73	9.11	[-27.95, 8.17]	14	-4.08	12.88	[-29.45, 21.42]	37	6.16	10.84	[-15.36, 27.13]	72
Feminine	-1.35	6.78	[-14.50, 12.07]	42	**-27.08**	**13.50**	**[-53.64, -0.55]**	**2**	6.88	9.93	[-12.32, 26.67]	76	6.59	8.95	[-11.08, 24.25]	77
first pass																
ORC	-32.69	33.50	[-98.14, 32.33]	16	2.05	14.21	[-26.49, 30.19]	56	-5.28	16.97	[-38.66, 27.73]	38	**-24.68**	**14.83**	**[-53.81, 4.57]**	**5**
Mismatch vs. Match	48.78	46.22	[-40.13, 140.73]	86	-36.29	37.50	[-108.98, 38.87]	16	20.94	22.70	[-22.88, 66.36]	82	29.35	20.49	[-10.20, 70.83]	93
Double vs. Single	**-111.34**	**48.84**	**[-207.91, -15.72]**	**1**	-21.41	15.98	[-53.09, 9.72]	9	-15.90	31.00	[-78.09, 45.32]	30	-13.55	24.57	[-62.77, 34.61]	29
Feminine	-16.85	32.28	[-80.04, 47.34]	30	-21.19	21.87	[-64.21, 22.96]	17	23.54	19.94	[-16.04, 62.08]	88	4.16	17.99	[-31.50, 39.54]	59
go-past																
ORC	-32.61	32.97	[-97.45, 32.04]	16	37.19	32.52	[-24.45, 104.24]	88	**84.57**	**52.12**	**[-17.34, 187.36]**	**95**	-32.12	75.15	[-181.03, 115.14]	33
Mismatch vs. Match	49.30	45.90	[-39.33, 142.78]	86	-115.29	74.77	[-261.64, 31.25]	6	-23.83	59.49	[-139.84, 92.58]	34	41.16	85.86	[-129.83, 210.34]	69
Double vs. Single	**-110.79**	**49.03**	**[-208.70, -15.14]**	**1**	-17.54	33.54	[-82.40, 49.79]	30	67.91	69.14	[-67.78, 203.46]	84	100.68	97.89	[-90.02, 295.24]	85
Feminine	-17.37	32.08	[-80.96, 45.57]	30	-49.10	40.91	[-128.68, 32.32]	11	**149.29**	**60.16**	**[31.30, 267.86]**	**99**	**143.64**	**83.19**	**[-20.48, 307.03]**	**96**
total duration																
ORC	51.35	62.71	[-73.14, 175.27]	80	-10.71	31.62	[-72.39, 50.51]	37	-20.14	38.08	[-94.44, 54.48]	29	-42.19	28.51	[-98.65, 14.16]	7
Mismatch vs. Match	64.59	85.65	[-103.10, 232.95]	78	**-382.77**	**103.12**	**[-581.19, -178.93]**	**0**	**91.71**	**49.09**	**[-3.98, 190.27]**	**97**	12.32	31.41	[-48.97, 74.11]	65
Double vs. Single	-77.30	80.37	[-233.17, 80.40]	17	37.12	44.07	[-49.01, 125.17]	80	-0.74	49.55	[-98.08, 96.47]	49	-7.09	39.14	[-84.14, 69.07]	43
Feminine	50.58	70.21	[-88.33, 188.36]	77	**-114.13**	**58.63**	**[-230.65, -0.32]**	**2**	**97.04**	**47.73**	**[1.21, 188.96]**	**98**	36.69	31.96	[-25.33, 100.07]	87

p(regress)																
ORC	-	-	-	-	0.18	0.18	[-0.16, 0.54]	85	0.08	0.16	[-0.24, 0.39]	70	0.16	0.20	[-0.27, 0.55]	79
Mismatch vs. Match	-	-	-	-	**-1.05**	**0.55**	**[-2.14, 0.02]**	**3**	-0.12	0.17	[-0.46, 0.22]	24	-0.04	0.19	[-0.41, 0.32]	41
Double vs. Single	-	-	-	-	0.28	0.25	[-0.21, 0.77]	87	**0.41**	**0.20**	**[0.01, 0.81]**	**98**	0.11	0.22	[-0.33, 0.54]	70
Feminine	-	-	-	-	-0.41	0.34	[-1.10, 0.24]	11	**0.31**	**0.16**	**[-0.01, 0.64]**	**97**	0.12	0.20	[-0.29, 0.51]	73
p(skip)																
ORC	-0.05	0.41	[-0.94, 0.68]	48	0.01	0.17	[-0.33, 0.33]	54	0.14	0.18	[-0.22, 0.48]	80	**0.33**	**0.18**	**[-0.01, 0.68]**	**97**
Mismatch vs. Match	-0.22	0.49	[-1.22, 0.74]	32	**0.90**	**0.50**	**[-0.05, 1.87]**	**97**	0.30	0.22	[-0.12, 0.75]	92	-0.04	0.18	[-0.41, 0.31]	41
Double vs. Single	0.48	0.48	[-0.47, 1.45]	85	0.06	0.22	[-0.37, 0.50]	61	0.14	0.24	[-0.32, 0.60]	72	-0.21	0.23	[-0.66, 0.23]	17
Feminine	0.47	0.33	[-0.18, 1.11]	93	**0.65**	**0.23**	**[0.20, 1.12]**	**100**	-0.16	0.21	[-0.56, 0.26]	22	-0.15	0.21	[-0.58, 0.26]	24

Table H.2: Linear and logistic mixed-effects model estimates of the dependent variables on all fixation metrics at spillover regions 2 and 3 for the Cor ORC vs. Cor SRC analysis, including SE estimates and CrIs. The % > 0 column shows the percent of the sampled posterior distribution that is over or under 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates that are bolded are significant.

Cor ORC vs. Cor SRC

	spillover 2				spillover 3			
Fixation Metrics	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**
first fixation								
ORC	**-18.37**	**9.20**	**[-36.64, -0.45]**	**2**	-14.98	9.68	[-34.22, 3.85]	6
Mismatch vs. Match	6.74	10.30	[-13.63, 26.81]	75	2.07	10.16	[-17.86, 22.25]	58
Double vs. Single	-13.04	10.76	[-34.23, 8.25]	11	-4.89	14.27	[-32.68, 23.53]	36
Feminine	-4.36	9.81	[-23.59, 15.17]	32	3.35	10.51	[-17.18, 23.97]	63
first pass								
ORC	-17.03	15.39	[-47.31, 12.98]	13	-8.24	18.93	[-45.04, 28.87]	33
Mismatch vs. Match	7.84	18.17	[-27.24, 44.41]	67	16.16	25.28	[-31.88, 68.00]	74
Double vs. Single	**-36.21**	**21.83**	**[-80.04, 5.90]**	**4**	-21.14	32.27	[-86.12, 42.41]	25
Feminine	1.55	16.74	[-31.68, 34.61]	53	**-44.62**	**22.87**	**[-90.54, 0.06]**	**3**
go-past								
ORC	-37.03	109.63	[-252.53, 172.95]	37	2.24	125.70	[-246.35, 251.18]	51
Mismatch vs. Match	-15.44	102.91	[-216.49, 189.30]	44	88.03	134.65	[-176.31, 350.86]	74
Double vs. Single	-136.35	112.53	[-360.36, 83.94]	11	-184.79	155.03	[-488.75, 119.71]	11
Feminine	-146.16	102.46	[-348.39, 56.09]	8	-20.34	137.19	[-292.26, 245.78]	44
total duration								
ORC	1.22	28.02	[-53.78, 56.23]	52	15.93	33.95	[-49.91, 82.51]	68
Mismatch vs. Match	23.03	33.28	[-42.36, 88.08]	75	61.72	40.69	[-15.40, 144.20]	94
Double vs. Single	-56.94	39.89	[-136.93, 20.02]	7	-18.87	39.80	[-97.74, 58.34]	31
Feminine	-4.36	30.43	[-64.41, 56.67]	44	-40.82	35.04	[-111.08, 27.87]	12
p(regress)								
ORC	0.22	0.22	[-0.22, 0.65]	84	0.47	0.31	[-0.13, 1.08]	94
Mismatch vs. Match	-0.14	0.20	[-0.54, 0.26]	24	0.25	0.33	[-0.41, 0.90]	78
Double vs. Single	-0.07	0.31	[-0.67, 0.54]	41	0.08	0.45	[-0.82, 0.95]	57
Feminine	-0.25	0.22	[-0.69, 0.18]	13	0.09	0.35	[-0.60, 0.78]	60
p(skip)								
ORC	0.19	0.17	[-0.15, 0.52]	87	0.12	0.17	[-0.21, 0.44]	78
Mismatch vs. Match	-0.08	0.16	[-0.41, 0.24]	30	-0.05	0.19	[-0.43, 0.31]	39
Double vs. Single	0.16	0.20	[-0.24, 0.54]	79	0.22	0.23	[-0.24, 0.67]	84
Feminine	0.22	0.18	[-0.14, 0.57]	89	0.07	0.17	[-0.27, 0.41]	65

Table H.3: Linear and logistic mixed-effects model estimates of the dependent variables on all fixation metrics at the matrix NP, RC pronoun, matrix verb, and spillover region 1 for the Incor ORC vs. Cor ORC analysis, including SE estimates and CrIs. The matrix NP region does not have an estimate for *p(regress)* as it is the first region in each sentence and a regressive saccade would be impossible. The % > 0 column shows the percent of the sampled posterior distribution that is over or under 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates that are bolded are significant.

Incor ORC vs. Cor ORC

	matrix NP				**RC pronoun**				**matrix verb**				**spillover 1**			
Fixation Metrics	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**
first fixation																
Incorrect	12.44	11.17	[-9.74, 34.21]	87	-15.89	17.99	[-51.41, 19.46]	19	2.27	16.07	[-29.33, 34.09]	56	11.33	14.11	[-16.00, 39.24]	79
Mismatch vs. Match	6.96	8.73	[-10.11, 24.14]	79	-30.69	29.76	[-88.48, 28.45]	15	4.45	14.31	[-23.79, 32.35]	62	2.04	15.78	[-29.12, 33.29]	56
Double vs. Single	2.46	9.86	[-16.81, 21.92]	60	-10.15	12.75	[-35.28, 15.01]	21	-13.98	16.20	[-46.16, 17.84]	19	-4.73	14.47	[-32.85, 23.99]	37
Feminine	-5.61	8.17	[-21.57, 10.46]	24	-24.85	16.72	[-57.51, 8.12]	7	4.51	13.64	[-22.31, 31.64]	63	9.83	12.00	[-13.79, 33.54]	80
first pass																
Incorrect	116.50	82.48	[-45.65, 279.41]	93	-22.73	26.86	[-76.05, 30.22]	20	20.06	34.43	[-47.52, 88.38]	72	21.94	29.57	[-35.42, 80.81]	77
Mismatch vs. Match	60.05	49.38	[-35.62, 156.96]	89	-62.97	52.25	[-164.88, 39.30]	11	-7.61	27.43	[-61.31, 46.56]	39	10.59	24.48	[-36.84, 58.95]	67
Double vs. Single	**-190.24**	**69.77**	**[-327.03, -51.17]**	**0**	-9.25	22.34	[-53.29, 34.89]	34	17.49	36.27	[-52.63, 89.68]	69	-14.77	23.66	[-61.84, 31.80]	26
Feminine	-20.19	41.75	[-102.32, 61.71]	31	-35.72	29.15	[-93.01, 21.36]	11	**45.42**	**27.71**	**[-9.22, 99.49]**	**95**	23.69	19.18	[-14.10, 61.52]	89
go-past																
Incorrect	116.78	81.89	[-43.01, 279.73]	93	-69.48	46.56	[-160.64, 21.44]	7	31.25	122.77	[-212.57, 271.25]	61	-34.88	109.80	[-249.84, 180.83]	37
Mismatch vs. Match	59.72	48.59	[-36.93, 155.65]	89	-19.59	103.47	[-223.82, 183.68]	42	-49.80	89.53	[-225.47, 125.42]	29	-47.71	90.76	[-224.97, 130.55]	30
Double vs. Single	**-189.39**	**69.55**	**[-326.17, -52.58]**	**0**	-71.84	58.62	[-186.87, 44.19]	11	112.49	117.65	[-119.85, 344.15]	83	107.81	112.53	[-111.00, 330.36]	84
Feminine	-20.17	41.63	[-102.05, 60.73]	31	-59.17	53.95	[-165.54, 47.41]	13	**200.30**	**88.70**	**[25.73, 375.12]**	**99**	**247.11**	**102.93**	**[44.28, 450.28]**	**99**
total duration																
Incorrect	37.18	106.70	[-172.58, 245.96]	64	16.31	56.30	[-94.72, 125.66]	62	25.96	76.80	[-123.70, 181.66]	64	52.31	56.25	[-57.42, 163.06]	83
Mismatch vs. Match	55.10	120.86	[-185.01, 291.13]	68	**-235.70**	**121.79**	**[-476.15, 5.32]**	**3**	49.28	58.07	[-63.79, 164.47]	81	-39.10	39.73	[-118.02, 39.05]	16
Double vs. Single	**-231.90**	**122.35**	**[-478.53, 6.44]**	**3**	15.67	54.14	[-93.05, 122.27]	62	23.96	59.06	[-90.86, 140.27]	66	-1.47	49.06	[-97.30, 93.47]	49
Feminine	-43.59	97.05	[-234.40, 144.43]	33	-73.19	71.75	[-213.74, 64.84]	15	**118.30**	**50.63**	**[18.29, 218.08]**	**99**	**80.94**	**43.65**	**[-4.12, 167.71]**	**97**
p(regress)																
Incorrect	-	-	-	-	**-0.78**	**0.54**	**[-2.02, 0.11]**	**5**	-0.12	0.26	[-0.65, 0.38]	34	-0.21	0.30	[-0.81, 0.37]	24
Mismatch vs. Match	-	-	-	-	-0.47	0.76	[-1.97, 1.01]	27	-0.15	0.22	[-0.59, 0.27]	24	0.18	0.26	[-0.33, 0.68]	76
Double vs. Single	-	-	-	-	0.11	0.47	[-0.79, 1.08]	59	0.23	0.30	[-0.36, 0.82]	78	0.30	0.33	[-0.34, 0.95]	83
Feminine	-	-	-	-	-0.22	0.44	[-1.13, 0.61]	30	0.13	0.21	[-0.29, 0.55]	74	0.37	0.27	[-0.17, 0.89]	91
p(skip)																
Incorrect	-0.61	0.62	[-1.99, 0.46]	15	0.26	0.39	[-0.56, 1.00]	76	-0.06	0.37	[-0.86, 0.61]	45	**-0.69**	**0.34**	**[-1.42, -0.08]**	**1**
Mismatch vs. Match	-0.40	0.58	[-1.60, 0.71]	23	0.92	0.64	[-0.34, 2.20]	92	0.20	0.27	[-0.31, 0.74]	78	0.23	0.24	[-0.25, 0.71]	84
Double vs. Single	0.21	0.55	[-0.90, 1.28]	66	0.11	0.29	[-0.44, 0.69]	64	0.42	0.30	[-0.14, 1.03]	93	-0.10	0.29	[-0.68, 0.46]	37
Feminine	0.20	0.40	[-0.59, 0.96]	71	0.41	0.33	[-0.24, 1.06]	90	-0.09	0.27	[-0.63, 0.46]	36	-0.14	0.25	[-0.64, 0.33]	28

Table 7.4: Linear and logistic mixed-effects model estimates of the dependent variables on all fixation metrics at spillover regions 2 and 3 for the Incor ORC vs. Cor ORC analysis, including SE estimates and CrIs. The % > 0 column shows the percent of the sampled posterior distribution that is over or under 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates that are bolded are significant.

Incor ORC vs. Cor ORC

	spillover 2				spillover 3			
Fixation Metrics	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**
first fixation								
Incorrect	24.69	18.12	[-11.10, 59.89]	92	-0.10	17.47	[-35.35, 33.28]	51
Mismatch vs. Match	-0.10	13.02	[-25.70, 25.62]	49	8.87	13.07	[-16.65, 34.44]	75
Double vs. Single	-16.45	13.77	[-43.49, 10.42]	12	-6.31	15.99	[-37.81, 25.21]	35
Feminine	-7.95	12.74	[-33.09, 16.76]	26	-4.16	12.92	[-29.93, 21.05]	37
first pass								
Incorrect	19.49	33.77	[-46.36, 85.31]	72	4.99	40.23	[-75.89, 83.97]	56
Mismatch vs. Match	11.80	26.23	[-38.93, 65.67]	67	32.64	31.13	[-27.43, 94.69]	86
Double vs. Single	**-77.69**	**45.39**	**[-167.43, 11.40]**	**4**	**-57.00**	**31.80**	**[-119.34, 4.92]**	**4**
Feminine	10.52	22.74	[-33.98, 55.83]	68	-41.46	26.85	[-94.65, 10.81]	6
go-past								
Incorrect	190.44	207.23	[-220.29, 596.14]	83	-73.09	233.59	[-532.95, 376.02]	38
Mismatch vs. Match	63.57	146.96	[-221.27, 355.38]	67	59.98	181.25	[-297.53, 418.77]	63
Double vs. Single	-141.92	165.48	[-463.10, 184.22]	19	-220.51	209.67	[-628.71, 193.74]	14
Feminine	-88.09	133.59	[-353.53, 173.92]	25	53.83	175.01	[-292.12, 398.19]	62
total duration								
Incorrect	14.04	52.28	[-86.91, 117.37]	61	-7.92	62.87	[-131.44, 116.15]	45
Mismatch vs. Match	12.19	41.32	[-68.41, 92.75]	61	52.83	46.65	[-37.36, 144.70]	87
Double vs. Single	-81.59	61.35	[-203.35, 39.59]	9	-52.34	51.72	[-153.92, 48.18]	15
Feminine	-12.67	44.82	[-101.55, 75.52]	39	-44.03	42.67	[-128.20, 39.84]	15
p(regress)								
Incorrect	-0.05	0.33	[-0.71, 0.59]	43	0.10	0.69	[-1.27, 1.48]	56
Mismatch vs. Match	-0.15	0.25	[-0.65, 0.34]	27	-0.37	0.46	[-1.28, 0.51]	20
Double vs. Single	-0.04	0.32	[-0.68, 0.59]	45	-0.26	0.69	[-1.63, 1.08]	35
Feminine	0.05	0.27	[-0.47, 0.58]	58	0.27	0.49	[-0.66, 1.28]	71
p(skip)								
Incorrect	-0.41	0.40	[-1.28, 0.30]	14	-0.16	0.32	[-0.80, 0.46]	31
Mismatch vs. Match	-0.14	0.22	[-0.58, 0.30]	26	-0.26	0.22	[-0.70, 0.17]	12
Double vs. Single	0.29	0.28	[-0.25, 0.85]	86	**0.51**	**0.28**	**[-0.03, 1.09]**	**97**
Feminine	0.25	0.26	[-0.28, 0.76]	83	0.24	0.22	[-0.20, 0.67]	86

Table H.5: Linear and logistic mixed-effects model estimates of the dependent variables on all fixation metrics at the matrix NP, RC pronoun, matrix verb, and spillover region 1 for the Incor ORC vs. Cor SRC analysis, including SE estimates and CrIs. The matrix NP region does not have an estimate for *p(regress)* as it is the first region in each sentence and a regressive saccade would be impossible. The % > 0 column shows the percent of the sampled posterior distribution that is over or under 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates that are bolded are significant.

Incor ORC vs. Cor SRC

	matrix NP				**RC pronoun**				**matrix verb**				**spillover 1**			
Fixation Metrics	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**
first fixation																
ORC	9.38	10.89	[-12.52, 30.31]	81	-17.88	15.37	[-48.53, 11.69]	12	-5.48	15.52	[-35.78, 25.02]	36	-5.55	13.68	[-32.33, 21.62]	34
Mismatch vs. Match	4.27	8.17	[-11.81, 20.40]	70	-13.90	28.60	[-69.48, 42.29]	31	1.38	13.71	[-25.92, 28.05]	54	**17.56**	**10.69**	**[-3.50, 38.83]**	**95**
Double vs. Single	7.77	9.97	[-11.82, 27.45]	79	-8.26	12.29	[-32.55, 15.92]	25	-9.01	14.01	[-36.48, 18.65]	26	1.97	13.29	[-24.44, 28.07]	56
Feminine	2.47	7.62	[-12.46, 17.52]	63	-13.22	17.61	[-47.88, 21.51]	22	1.69	11.76	[-21.34, 25.07]	56	0.46	11.26	[-21.55, 22.48]	52
first pass																
ORC	40.54	81.54	[-115.57, 205.08]	69	-11.75	21.50	[-54.60, 30.39]	29	3.50	31.90	[-58.91, 66.75]	55	-18.95	30.31	[-78.32, 40.51]	27
Mismatch vs. Match	65.22	52.24	[-39.04, 167.82]	90	-62.47	47.12	[-154.60, 29.22]	9	17.06	28.67	[-38.91, 73.99]	73	34.61	22.66	[-9.81, 78.61]	94
Double vs. Single	-48.72	51.72	[-149.46, 53.75]	17	-13.29	20.76	[-54.17, 27.76]	26	-40.95	28.46	[-96.83, 14.30]	7	-15.22	32.49	[-80.81, 48.05]	32
Feminine	-10.44	42.97	[-95.46, 74.07]	40	-16.94	27.38	[-70.57, 37.29]	27	5.04	25.05	[-43.89, 54.35]	58	-5.78	22.00	[-49.30, 37.18]	40
go-past																
ORC	39.97	83.04	[-121.53, 205.94]	69	-26.05	46.35	[-120.77, 62.33]	29	82.76	136.78	[-191.70, 346.63]	74	-92.19	110.81	[-307.36, 127.89]	20
Mismatch vs. Match	65.59	52.82	[-40.31, 170.02]	90	**-188.00**	**84.65**	**[-353.66, -21.73]**	**1**	14.40	77.59	[-139.36, 167.08]	57	60.01	100.33	[-137.13, 259.91]	73
Double vs. Single	-48.18	52.02	[-149.40, 54.91]	18	-4.77	36.79	[-76.57, 68.18]	45	74.23	78.49	[-80.16, 228.72]	83	25.68	137.89	[-248.37, 296.34]	58
Feminine	-10.95	42.60	[-94.61, 73.20]	40	-34.28	56.29	[-144.50, 76.06]	27	**159.49**	**70.47**	**[20.78, 295.05]**	**99**	114.32	98.34	[-78.14, 311.85]	88
total duration																
ORC	21.26	102.90	[-178.07, 226.62]	58	32.34	52.09	[-70.13, 134.09]	73	-7.99	60.68	[-125.17, 112.44]	45	-10.63	56.02	[-120.26, 99.89]	42
Mismatch vs. Match	**167.74**	**96.25**	**[-22.72, 357.17]**	**96**	**-289.67**	**131.42**	**[-546.97, -30.48]**	**1**	95.37	63.28	[-28.95, 219.79]	94	26.81	36.80	[-45.15, 99.46]	77
Double vs. Single	-27.99	86.79	[-200.15, 144.24]	37	19.60	55.50	[-89.76, 129.13]	64	-34.99	56.26	[-144.86, 75.44]	26	-54.36	44.51	[-143.19, 32.78]	11
Feminine	**118.64**	**71.51**	**[-21.77, 258.38]**	**95**	-42.32	70.49	[-181.95, 95.49]	27	**133.19**	**52.49**	**[30.22, 235.85]**	**99**	18.18	36.31	[-52.70, 89.14]	70
p(regress)																
ORC	-	-	-	-	-0.52	0.53	[-1.74, 0.34]	15	0.07	0.25	[-0.44, 0.55]	62	-0.07	0.26	[-0.61, 0.43]	40
Mismatch vs. Match	-	-	-	-	**-2.18**	**0.74**	**[-3.65, -0.75]**	**0**	0.03	0.20	[-0.37, 0.43]	55	-0.24	0.21	[-0.65, 0.16]	12
Double vs. Single	-	-	-	-	0.08	0.33	[-0.58, 0.72]	60	**0.67**	**0.22**	**[0.23, 1.11]**	**100**	-0.31	0.30	[-0.92, 0.26]	15
Feminine	-	-	-	-	**-1.38**	**0.55**	**[-2.51, -0.34]**	**0**	**0.53**	**0.19**	**[0.15, 0.90]**	**100**	0.02	0.22	[-0.41, 0.45]	55
p(skip)																
ORC	-0.37	0.67	[-1.95, 0.69]	31	0.16	0.36	[-0.62, 0.82]	71	0.09	0.35	[-0.65, 0.72]	63	-0.23	0.35	[-0.97, 0.41]	26
Mismatch vs. Match	0.15	0.54	[-0.92, 1.23]	61	0.72	0.63	[-0.52, 1.94]	87	0.34	0.24	[-0.13, 0.82]	92	-0.04	0.22	[-0.46, 0.39]	44
Double vs. Single	0.45	0.52	[-0.58, 1.50]	82	0.04	0.29	[-0.53, 0.63]	56	0.35	0.26	[-0.17, 0.87]	91	-0.28	0.28	[-0.83, 0.27]	16
Feminine	0.50	0.42	[-0.33, 1.32]	89	**0.66**	**0.31**	**[0.06, 1.27]**	**98**	0.07	0.23	[-0.37, 0.53]	62	-0.14	0.28	[-0.69, 0.41]	30

Table H.6: Linear and logistic mixed-effects model estimates of the dependent variables on all fixation metrics at spillover regions 2 and 3 for the Incor ORC vs. Cor SRC analysis, including SE estimates and CrIs. The % > 0 column shows the percent of the sampled posterior distribution that is over or under 0; values that are >= 95 for positive estimates or <= 5 for negative estimates are considered significant. Estimates that are bolded are significant.

Incor ORC vs. Cor SRC

	spillover 2				spillover 3			
Fixation Metrics	**β**	***SE***	**CrI**	**%>0**	**β**	***SE***	**CrI**	**%>0**
first fixation								
ORC	0.69	15.62	[-30.17, 31.48]	52	-15.42	16.67	[-48.90, 17.21]	17
Mismatch vs. Match	7.69	13.93	[-19.56, 35.31]	71	-5.13	12.62	[-29.76, 20.02]	34
Double vs. Single	-9.30	15.23	[-39.20, 20.90]	27	-2.17	17.72	[-36.99, 32.56]	45
Feminine	-15.98	11.42	[-38.46, 6.26]	8	12.81	14.66	[-16.19, 41.16]	81
first pass								
ORC	8.26	31.51	[-54.14, 70.32]	61	-16.38	41.68	[-100.10, 64.67]	35
Mismatch vs. Match	26.72	22.08	[-16.26, 70.44]	89	15.56	27.49	[-38.35, 69.60]	72
Double vs. Single	-38.98	32.75	[-104.13, 25.90]	11	-26.59	38.62	[-101.93, 50.48]	24
Feminine	4.38	24.90	[-44.11, 53.79]	57	-43.53	27.83	[-98.45, 11.73]	6
go-past								
ORC	125.73	162.88	[-188.78, 449.50]	78	-77.79	210.03	[-490.96, 328.81]	36
Mismatch vs. Match	92.36	115.73	[-133.92, 318.23]	79	205.38	170.12	[-126.14, 534.29]	88
Double vs. Single	**-346.79**	**142.57**	**[-624.18, -65.22]**	**1**	-134.10	185.23	[-499.12, 230.51]	23
Feminine	-68.34	112.60	[-289.97, 152.29]	27	-139.45	155.66	[-448.09, 165.76]	19
total duration								
ORC	10.19	47.19	[-82.44, 102.56]	58	-13.98	79.40	[-172.58, 142.12]	43
Mismatch vs. Match	32.12	35.88	[-37.14, 103.75]	82	34.66	42.20	[-48.26, 118.01]	80
Double vs. Single	-75.38	51.34	[-176.01, 25.78]	7	-47.68	51.10	[-148.90, 52.34]	17
Feminine	18.90	38.60	[-56.30, 95.06]	69	-27.09	39.38	[-103.96, 49.91]	24
p(regress)								
ORC	0.25	0.33	[-0.41, 0.91]	78	0.15	0.63	[-1.07, 1.43]	60
Mismatch vs. Match	0.05	0.26	[-0.46, 0.56]	58	0.40	0.40	[-0.37, 1.23]	85
Double vs. Single	-0.32	0.34	[-0.99, 0.35]	17	-0.13	0.55	[-1.23, 0.97]	40
Feminine	-0.22	0.29	[-0.78, 0.35]	22	0.19	0.41	[-0.62, 0.99]	68
p(skip)								
ORC	-0.17	0.38	[-1.00, 0.50]	35	-0.04	0.33	[-0.71, 0.59]	47
Mismatch vs. Match	-0.08	0.21	[-0.49, 0.33]	35	0.27	0.23	[-0.18, 0.72]	88
Double vs. Single	0.13	0.25	[-0.36, 0.61]	70	-0.04	0.27	[-0.57, 0.47]	44
Feminine	**0.33**	**0.20**	**[-0.08, 0.73]**	**95**	-0.22	0.22	[-0.66, 0.21]	15