Acceptability & the Ambiguity Advantage Effect as a Window into Parsing Behavior

Christian Jorge Muxica

University of Massachusetts Amherst, Department of Linguistics

Abstract

A long standing question within the field of sentence processing is the extent to which the parser is able to maintain multiple representations of ambiguous input during the act of comprehension. There are two broad classes of sentence processing models defined by the answer to this question- serial and parallel. In serial models only one representation of the input can be maintained. While multiple syntactic descriptions of the input are utilized by parallel models. In the first experiment, we address this question by modifying the speeded acceptability judgement paradigm utilized by Dillon et al. (2019). We make these modifications with the aim of providing better controlled data, suitable for the Drift Diffusion model (Ratcliff, 1978). In the second experiment, we continue with this paradigm and manipulate our stimuli to delay disambiguation. This is done to evaluate the Gradient Symbolic Computation parser (Cho et al., 2017), specifically the predicted transition from parallel-like continuous representations to discrete serial-like ones. With this combination of behavioral data and computational analysis we intend to asses the validity of speeded acceptability judgements as a sentence processing paradigm, address the fundamental question of serial versus parallel processing, and evaluate a model which rejects this architectural binary.

Keywords: sentence processing, speeded acceptability-judgements, drift diffusion, parallelism, gradient symbolic computation

1. Introduction

An enduring question in sentence processing is the extent to which the parser is able to maintain multiple representations of ambiguous input. There are two broad classes of sentence processing model defined by this question-

serial and parallel. In a serial model, only one representation of the input can be maintained at any given point. Thus, the introduction of structural ambiguity constitutes a fork in the syntactic road where a single parse must be chosen. This decision might be made probabilistically or using some structural criterion. In a parallel model, multiple syntactic descriptions of the input are generated, maintained, and manipulated over comprehension. Parsing decisions are still involved in parallel processing, but these take a different form such as the choice to shift the relative weights of different parses. However, unlike a serial model, these operations are not immediately required to resolve to a single analysis.

Parallelism is a coarse distinction. Models can still vary with regard to factors such as working memory constraints, the nature of representations, or reanalysis strategies. While, undoubtedly, each of these factors has empirical implications for comprehension, these considerations exist at a finer level of implementation than we intend to address. Rather, the aim of this research is to answer the aforementioned question in the broadest implementational sense: does the parser maintain multiple representations of ambiguous input during sentence comprehension? To investigate this question, we present results from a speeded acceptability experiment, modified from the paradigm developed by Dillon et al. (2019), and subsequent analysis utilizing the Drift Diffusion Model of binary choice tasks developed by Ratcliff (1978).

2. The Ambiguity Advantage Effect

Syntactic ambiguities can take one of two forms- global or local. Ambiguities which are global lack any definitive cues to force syntactic resolution. For instance, the first sentence in Example 1 has two possible interpretations. One in which "the boy" is using a telescope to see "the girl" and a second in which "the boy" is seeing "the girl" who is in possession of a telescope. Preferred readings will vary across languages and individuals, but no material from within the sentence rules out either analysis. Local ambiguities in the other hand resolve over the course of a sentence. The same ambiguity from the first sentence of Example 1 is also present in the second, but only temporarily. The reading in which the boy holds the telescope is disconfirmed at the prepositional phrase "in her hand" as the feminine pronoun agrees with the girl. Structural ambiguity is often a key manipulation in psycholinguistic experimental design, by and large with the intended effect of yielding

processing difficulty. In spite of this, there are instances in which sentence comprehension seems to excel on account of structural ambiguity.

(1) Global versus Local Structural Ambiguity
   a. The boy saw the girl with the telescope.
   b. The boy saw the girl with the telescope in her hand.

The ambiguity advantage effect is the puzzling finding that globally ambiguous sentences are processed more quickly than lexically matched yet unambiguous counterparts. Traxler et al. (1998) originally observe this effect in relative clause attachment ambiguities within an eyetracking while-reading experiment. As seen in Example 2 the ambiguity was manipulated by virtue of plausibility- cars cannot have facial hair. The authors find that total reading times for the region corresponding to "a moustache" were lower on average for the ambiguous condition as compared to either unambiguous condition. Traxler et al. (1998) take this counterintuitive effect as evidence in favor of the serial unrestricted race model. This model synthesizes aspects of contemporary serial two-stage models (Frazier, 1979, 1987) and parallel constraint-based models (MacDonald, 1994; McRae et al., 1998). The Unrestricted Race Model (URM) is similar to two-stage models in that it is serial, a single parse is predicted and reanalysis takes place should that parse become disconfirmed. What the URM takes from constraint-based models is the claim that any kind of information (structural, semantic, discourse, or otherwise) can influence early decisions in structure building. Under the URM, when the parser arrives at the complementizer a relative clause must be posited. However, the preceding material provides no cue as to where this relative clause should be attached. Thus an attachment site, either high or low, must be chosen by the URM for the single representation under construction. In the ambiguous condition, either attachment will result in a plausible sentence at disambiguation. However, in the unambiguous conditions, the plausibility at disambiguation will depend on the attachment chosen. Take the high attachment condition, if the parser has chosen the incorrect low attachment at the complementizer, processing will continue as normal until "a moustache" is reached. Here, the low parse will become implausible and difficulty will be experienced as reanalysis takes place to find the correct analysis. If the parser has chosen the correct high attachment, processing will continue uninterrupted through the entire sentence. Under this analysis, the increased average reading times in the unambiguous conditions are the

$$difficulty \quad \alpha \quad -log(P(w_i|w_{1...i-1}))$$

Figure 1: Processing difficulty predicted at a given word by surprisal from Levy (2008)

result of the subset of trials where the incorrect parse is chosen. This predicts that participant behavior on unambiguous trials will break down into two distinct groups. One where the incorrect parse is chosen and difficulty is experienced and another where the correct parse is chosen and no difficulty is experienced.

(2) Three conditions from Traxler, Pickering, & Clifton (1998)
   a. Ambiguous: The son of the driver that had a moustache was really cool.
   b. High: The driver of the car that had a moustache was really cool.
   c. Low: The car of the driver that had a moustache was really cool.

While the ambiguity advantaged effect is incompatible with constraint-based models, there are parallel models for which this result is not a concern. Imagine a generalized parallel model in which processing difficulty is defined using Levy (2008)'s surprisal theory. Under this proposal, the processing difficulty experienced at the current word (w_i) is equivalent to surprisal: the negative log probability of a word conditioned on the previous material ($w_{1...i-1}$) within an incremental context. As a word becomes more probable within a context, the surprisal and predicted processing difficulty for that word decrease. The reverse is true when words are less contextually probable-surprisal and processing difficulty increase. Surprisal theory abstracts away from implementational detail and is compatible with a variety of sentence processing models as a result.

Returning to the conditions from Traxler et al. (1998), when this model arrives at the complementizer it will posit a relative clause, representing both attachment sites in some capacity. The processing difficulty predicted across conditions will be proportional to the surprisal of the disambiguating region. This surprisal value will be a function of the conditional probability of "a moustache" within the preceding context. Due to the fact that the model is parallel, the conditional probability will be a weighted sum given both the high and low parse.

In the ambiguous condition, both parses attach the relative clause to an animate argument which could plausibly have a moustache. The resulting

$$P(moustache) = P(RC_{low})P(moustache|RC_{low}) + P(RC_{high})P(moustache|RC_{high})$$

Figure 2: Surprisal for critical word in Traxler et al. (1998) items

summed conditional probabilities will be relatively high, leading to low surprisal at disambiguation. In the unambiguous conditions only one parse will make a plausible attachment. Thus, the joint conditional probability will be relatively low, leading to higher surprisal. The inverse relationship between surprisal and processing difficulty correctly predicts the pattern that Traxler et al. (1998) observe. Ambiguous material is easier to process than unambiguous material.

While the URM and suprisal both account for the ambiguity advantage effect, these models predict different distributions of processing difficulty for the unambiguous conditions. In every trial surprisal is a combination of a plausible and implausible parse, thus reading times should be stable and consistently higher than the ambiguous condition- a unimodal distribution. The unrestricted race model predicts a bimodal distribution. On a subset of trials, the correct parse will be chosen and reading times will be low and nearly identical to the ambiguous condition. On another subset of trials, the incorrect parse will be chosen and reading times will be high and distinct from the ambiguous condition. The pattern of inter-trial behaviour observed under local ambiguity would test these predictions.

3. Eyetracking & Modal Distributions

Intuitively, the response pattern modality for the unambiguous conditions should be present in standard reading measures. Under the URM we might expect high variation in reading times or the probability of regression at disambiguation, while a parallel model might predict a more stable pattern of difficulty. Unfortunately intuition conflicts with a number of methodological obstacles. Chiefly, reading experiments do not provide a sufficient amount of data for the accurate estimation of modes (Gibson and Pearlmutter, 2000). Thousands of observations would be required and a typical eyetracking experiment offers only hundreds. Further, Gibson and Pearlmutter (2000) argue that the distribution of the individual modes predicted by a serial model would be difficult to tease apart from one another. If one mode is larger than the other, indicating that it is observed more often, then by default it would overlap more with the second mode. This would give the illusion of a

more homogeneous distribution. As well, if one mode is substantially distant from the other, indicating that difference in processing difficulty observed is drastic, then by default it must be smaller than the other distribution. This would make it difficult to distinguish a small distant mode from the dissipating tail of a unimodal distribution (Gibson and Pearlmutter, 2000). For all of these reasons, a typical reading experiment would likely obfuscate an underlying bimodal response distribution should it even exist. Thus, we turn to acceptability.

4. Acceptability & Incremental Parsing

End-of-sentence acceptability responses, both speeded and unspeeded, reflect sentence medial processing difficulty (Tabor and Hutchins, 2004; Clifton and Frazier, 1998; Ferreira and Henderson, 1991). From this it follows that the ambiguity advantage effect should be present in acceptability judgements. The processing difficulty associated with unambiguous sentences should be reflected in lower perceived acceptability and increased reaction times to provide judgements. The reverse should then be true for easier globally ambiguous sentences. Additionally, acceptability experiments address the concerns that Gibson and Pearlmutter (2000) raise. Bimodal response behaviour has been observed in acceptability (Dillon et al., 2017) and localizing the dependent measure to a single judgement affords more observations than a typical reading experiment. These facts indicate that acceptability would be a suitable dependent measure for the investigation of parallelism.

(3) Four conditions from Dillon, Wagers, Andrews, & Rotello (2019)
   a. NoMatch: Armand spotted the cousins of the painters who knits.
   b. MultiMatch: Armand spotted the cousin of the painter who knits.
   c. LowMatch: Armand spotted the cousins of the painter who knits.
   d. HighMatch: Armand spotted the cousin of the painters who knits.

In search of modality in response behaviour, Dillon et al. (2019) attempt to replicate the ambiguity advantage effect within an end-of-sentence speeded acceptability experiment. Stimuli were shown to participants in rapid serial visual presentation- one word at a time in the center of a computer screen. After the final word in the sentence, subjects were given a speeded judgement task followed by an unspeeded three point confidence rating. As in Traxler

et al. (1998) relative clause ambiguities are utilized, but with three significant changes. First, the ambiguity was manipulated by virtue of number agreement between subject and verb rather than plausibility. Second, the locus of disambiguation, the verb, was generally pushed toward the end of the sentence in an attempt to reduce the potential for reanalysis. Lastly, a NoMatch condition was added to provide a baseline of ungrammatical response behaviour for analysis.

		"Acceptable"	**"Unacceptable"**
Condition	**Accuracy**	**Reaction Time (Standard Error)**	
NoMatch	84%	1531 (87)	1185 (35)
MultiMatch	74%	1267 (34)	1458 (59)
LowMatch	60%	1356 (41)	1464 (52)
HighMatch	41%	1455 (49)	1352 (41)

Table 1: Average reaction times in milliseconds and accuracy of judgements across all four conditions in Dillon et al. (2019), correct responses highlighted in grey

The results from Dillon et al. (2019) reproduce the ambiguity advantage effect. Participants accept sentences in the MultiMatch condition more often than sentences in the LowMatch and HighMatch conditions. Further, reaction times to accept the MultiMatch condition were on average lower than reaction times to accept either the LowMatch or HighMatch conditions. Beyond replicating the ambiguity advantage effect, Dillon et al. (2019) observe strong evidence of parallel parsing strategies. The pattern of lower confidence and slower reaction times to judge unambiguous sentences as acceptable receives a simple explanation under a parallel model. The co-activation of both the grammatical and ungrammatical attachments at, and shortly after, the point of disambiguation creates conflicting signals about the acceptability of the sentence. This could produce the consistent difficulty experienced by participants on these unambiguous conditions. A serial model on the other hand offers no explanation for this kind of effect. Assuming that a full reanalysis was not possible, it is unclear how the chosen grammatical attachment in these cases would be influenced by the unchosen ungrammatical attachment.

While the results of Dillon et al. (2019) support a parallel parser, this conclusion hinges upon two critical assumptions. First, there exist is a strong relationship between response and the current internal representations. Other possibilities exist. Imagine a serial model under which the perceived acceptability of a sentence reflects the average acceptability of all parses pursued.

Given time to reanalyze, this style of model would allow both high and low attachments to influence the judgement task, creating an illusion of co-activation. For the unambiguous conditions the average acceptability would comprise one acceptable and one unacceptable parse. This could create a weaker sense of acceptability and fit the lower accuracies and slower reaction times in these conditions. Out of the forty critical items that Dillon et al. (2019) utilize, the verb occurred as the final word for nineteen items (47.5%), as the penultimate word in sixteen items (40%), and was followed by two words in the remaining five items (12.5%). The reanalysis process would have to be highly rapid under such a model. Take the 40% of trials where the verb occurred in the penultimate position. Each word was presented for 225ms and followed by 100ms of a blank screen. Thus, there exists a 650ms window between disambiguation and the judgement task in which to reanalyze. This is not an insignificant amount of time, but it is a small interval to reanalyze a sentence within. The extent to which this is possible within such a parse-averaging model remains an open question. The second fact upon which the conclusions of Dillon et al. (2019) depend is the interpretability of acceptability as an experimental measure. "Acceptable" and "unacceptable" are not precise terms and participants require definitions within the instruction phase. The incremental process a participant takes to arrive at one of these categories is in no way transparent. As well, this paradigm does not control for the possibility of response bias which could dramatically impact results. The current design intends to address these concerns by means of design manipulations and computational modelling.

5. The Drift Diffusion Model

First described in Ratcliff (1978), the drift diffusion model is a model of the incremental process of arriving at a choice within a binary decision task given some information. It is is entirely agnostic to the nature of the information and the two options involved in the decision task- allowing the model to apply to a wide range of phenomena.

Depicted in Figure 3 is the drift diffusion model as applied to a grammaticality judgement. In drift diffusion, a binary decision task is modeled as an incremental process of "noisy" evidence accumulation. Each of the meandering lines in the figure correspond to the time course of this evidence accumulation. The top and bottom lines in the figure represent thresholds that correspond to choices in the decision task. When evidence accumula-

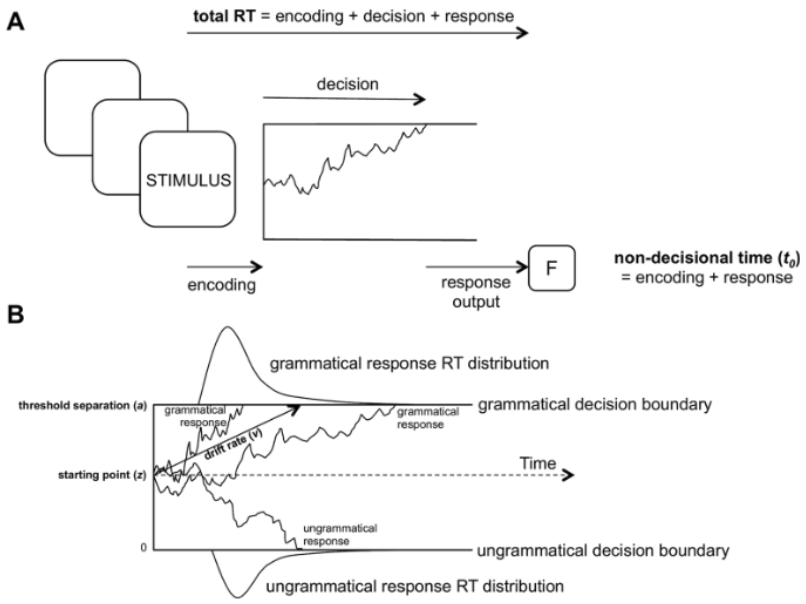

Figure 3: Schema of the drift diffusion model as applied to acceptability judgements, taken from Hammerly et al. (2019)

tion surpasses a threshold, the respective choice has been made and motor response begins. The starting point for evidence accumulation is specified by the parameter z which can be thought of as bias. The closer z is to some decision thresholds, the stronger bias is toward that decision. Threshold separation or a describes the distance between thresholds and the amount of information required to arrive at a decision. Lastly, the rate of evidence accumulation is specified by v and is referred to as drift rate. This describes the quality of the information extracted from the stimulus- higher quality information creates a higher rate of evidence accumulation. In the present experiment, drift rate would correspond to the strength of perceived acceptability created by the current parse(s) at judgement. The drift diffusion model is "noisy" in that the path toward a decision is not a straight line with a slope dictated by the drift rate. Rather, evidence accumulation is stochastic in nature, meaning with some chance the point will drift away from the threshold which the information pulls it toward. This accounts for the possibility of errors in judgement. There are a variety of other free parameters which make up a full drift diffusion model, but the three described above are most relevant for the present study.

The value of this model is in the credence it grants to the interpretation of results. As outlined earlier, acceptability is opaque as a dependent measure of incremental parsing. The drift diffusion model captures the internal mechanics of the decision-making process in a way that accuracy and reaction times alone cannot. Conducting an estimation of the parameters will quantify confounding factors and critical information such as evidence quality by condition. As well, should this established model of decision-making achieve a strong fit to the results, then speeded acceptability will be receive some support as a stable measure of incremental sentence processing.

6. Experiment 1

There are two objectives with the design of the first experiment. First, to replicate the ambiguity advantage effect within speeded acceptability judgements elicited at disambiguation. Second, to evaluate the modal predictions of serial and parallel parsing models. Subsequently, these results are fit using the drift diffusion model to control for response bias and provide additional insights into the results. Critical items are written in this experiment such that disambiguation and the judgement task occur simultaneously. This eliminates the potential for rapid serial reanalysis, providing better controlled stimuli for the evaluation of response modality and the use of the drift diffusion model.

6.1. Method

6.1.1. Participants

Forty eight participants were recruited from an advertisement posted to Prolific, a participant recruitment site similar to Amazon's MechanicalTurk. All participants were American, monolingual native English speakers, between the ages of eighteen and fifty, with no known language related disorders. Participants were compensated with $5.00 U.S. dollars. This procedure was approved by the University of Massachusetts Amherst Internal Review Board.

6.1.2. Materials

All forty of the critical experimental items were taken from Dillon et al. (2019) and manipulated such that the final word presented to participants was the disambiguating verb. This adjustment was made to eliminate the possibility of rapid serial reanalysis, as this was a concern with around half

of the items Dillon et al. (2019) utilize. Of the forty critical items taken and adjusted, nineteen (47.5%) ended on a non-copula verb, meaning participants were providing an acceptability judgement on a feasibly complete sentence. The remaining twenty-one items (52.5%) however, ended on the past tense form of the copula as Example 4 shows. On these trials participants provided acceptability judgements on an incomplete sentence. Participants were warned of this fact before the experiment and were instructed to judge the quality of fragments based on the quality of the material leading up to and including the final word. These forty items were Latin Squared into four distinct lists and each of these lists was combined with the same set of sixty-eight filler items for each participant.

(4) Four conditions from experiment 1
   a. NoMatch: Karl recognized the hostages of the pirates who was
   b. MultiMatch: Karl recognized the hostage of the pirate who was
   c. LowMatch: Karl recognized the hostages of the pirate who was
   d. HighMatch: Karl recognized the hostage of the pirates who was

The filler items were taken from Dillon et al. (2019) and manipulated such that the acceptability judgement was elicited at a verb. This was done to mirror the critical items. Forty-four of the sixty-eight fillers were ungrammatical, which combined with the ten ungrammatical critical items ensures each participant was exposed to an even split of grammatical and ungrammatical stimuli. The ungrammaticality of these forty-four filler items came from a variety of sources such as agreement with conjoined subjects, verb plausibility (e.g. ...her assistant shattered.), and various morphosyntactic errors (ex. ...finally able to performed). These fillers were included to reduce the potential of scanning strategies for number agreement on the critical items.

6.1.3. Procedure

Participants were directed from Prolific to an experiment hosted on Ibex Farm. The paradigm was described in an instruction phase and definitions for acceptability (specifically with regard to colloquial English) were provided. Additionally there was a practice period with five test items. At the end of the experiment, participants were debriefed with questions about the experience with the paradigm and an open ended question utilized to discriminate bots from humans.

Stimuli were presented in rapid serial visual presentation. Each trial in the experiment began with a fixation cross displayed for 1 second in the center of the screen. Immediately following, each word in the sentence (or sentence fragment) would be displayed for 225 milliseconds in the center of the screen with 100 milliseconds of a blank screen intervening. The final word in each trial was presented in a green font. This indicated to participants that the word currently displayed was the final word and the judgement task had begun. This technique developed by Hammerly et al. (2019) removes any window between disambiguation and the judgement task. Participants were instructed to provide this acceptability judgement as quickly as possible while maintaining accuracy. Responses were recorded using two keys on the keyboard, 'f' indicating that item was acceptable and 'j' indicating that the item was unacceptable. The experiment took around 25-35 minutes for participants to complete.

6.1.4. Fast-DM

To calculate drift diffusion parameter estimations we utilize the program fast-dm (Voss and Voss, 2007) specifically version 30.2 (Voss et al., 2015). Fast-dm takes the response data from each participant in an experiment and fits an individual diffusion model for each one. Thus, the output of fast-dm is a set of parameter estimations for each participants. Inferential statistics are then performed on this set of values. This is standard practice in diffusion modeling (Voss et al., 2013).

Fast-dm version 30.2 offers three possible optimization criteria for the estimation of these parameters- maximum-likelihood, Kolmogorov-Smirnov, and chi square. We use the Kolmogorov-Smirnov (KS) approach as it is a compromise between the stability of the chi square criterion and the efficiency of maximum-likelihood estimation (Voss et al., 2015). We provide a brief overview of the KS approach as it is valuable for the evaluation of model fit.

The KS optimization method for drift diffusion modelling (Voss and Voss, 2007) is based on the statistic of the Kolmogorov-Smirnov test (Kolmogoroff, 1941). The KS statistic quantifies the maximum vertical distance between a predicted and empirical cumulative distribution function- here a distribution of reaction times. The KS optimization method works by finding the parameter set (and corresponding predicted reaction time distribution) which minimizes the value of this statistic. The drift diffusion model outputs two separate reaction time distributions, one for decisions made at the upper threshold and another for the lower threshold. For convenience, reactions

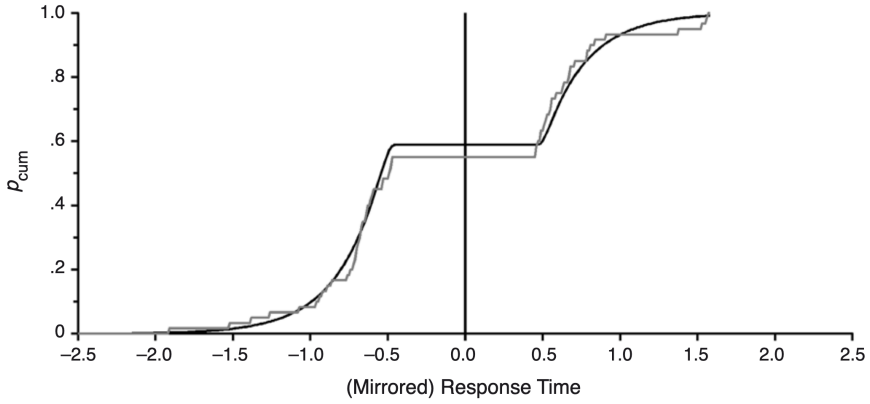

Figure 4: Example comparison of empirical and predicted response time distributions. The ascending black line is the accumulated probability function computed according to the diffusion model. The gray line shows the cumulative probability of empirical response times. Figure taken from Voss et al. (2004).

times for the lower threshold are multiplied by -1 yielding a single continuous distribution function. The cumulative distribution function (CDF) returns the probability of observing reaction times lower than or equal to a given reaction time. This is visualized in Figure 4.

At the end of optimization, a p-value calculated from the KS statistic of the final predicted CDF is returned as a measure of fit. Values of p below 0.05 indicate that the diffusion model fitted does not account for the data observed. However, this measure of fit can be unreliable. When one or more parameters is free to vary between conditions (as drift rate is for our models) fast-dm calculates the product of all p-values from the different conditions. This can have the unintended consequence of yielding low p-values (Voss et al., 2013). As well, in the case of small trial numbers (as is the case for our 40 trial experiment) the power of the KS test is smaller and ill fitting estimations might not be detected (Voss et al., 2013). Thus, it is critical when evaluating model fit to include graphical displays of the relationship between the predicted and observed distribution of reaction times. We present these figures in our results.

6.2. Results

Of the 48 participants recruited 39 were utilized in the analysis of results. Two of these participants were rejected for having response accuracy on the

filler items 2.5 standard deviations below the mean. The remaining 7 were rejected for having one or more judgement reaction times above 6 seconds for the critical items. Participants were instructed to provide their judgements as fast as possible while maintaining accuracy. These trials indicated that the participant was not performing the task as instructed. Simply removing individual trials would have left too few observations for conditional drift diffusion parameter estimation, thus these participants were excluded from analysis entirely.

6.2.1. Accuracy

Looking at judgement accuracy across conditions, we observe that participants have the strongest performance in the ambiguous NoMatch and MultiMatch conditions. Overwhelmingly participants correctly judge the NoMatch condition to be unacceptable and the MultiMatch conditions to be acceptable. The results are more mixed for the unambiguous conditions. In the LowMatch condition participants approach accuracy that is comparable, but lower than that of the NoMatch condition. There is a dramatic drop in accuracy for the HighMatch condition where accuracy approaches chance.

6.2.2. Reaction Time

Turning to reaction times for correct responses, we observe the lowest reaction times, by a substantial margin, for the MultiMatch condition. The variation in reaction times as characterized by the standard error is the lowest in this cases as well. The second lowest reaction time and standard error occur for the LowMatch condition. Correct NoMatch response were provided notably slower than both the MultiMatch and LowMatch conditions, but standard error is close among the three. We observe the weakest performance for the HighMatch condition. Here reaction times and variation are both comparatively high. Judgements are provided on average 400ms slower than the MultiMatch condition for example and standard error is around double that of the other three conditions.

For the NoMatch, MultiMatch, and LowMatch conditions we find that reaction times and standard error are higher for incorrect responses. This generalization does not hold for the HighMatch condition. Here, the incorrect response is provided faster, more often, and with less variation than the correct response.

		"Acceptable"	"Unacceptable"
Condition	**Accuracy**	**Reaction Time (Standard Error)**	
NoMatch	79%	1511 (78)	1357 (39)
MultiMatch	88%	1145 (31)	1516 (139)
LowMatch	73%	1252 (37)	1428 (82)
HighMatch	48%	1557 (70)	1355 (52)

Table 2: Average reaction times in milliseconds and accuracy of judgements across all four conditions in Experiment 1, correct responses highlighted in grey

Parameter	**Mean**	**SE**
a	2.002	0.077
zr	0.509	0.022
vNoMatch	-0.990	0.118
vMultiMatch	1.654	0.209
vLowMatch	0.819	0.154
vHighMatch	-0.157	0.127

Table 3: Mean values and standard error of estimated drift diffusion parameters in experiment 1, drift rates are estimated by condition

6.2.3. Drift Diffusion Parameter Estimations

As discussed, fast-dm requires that one decision threshold be coded as negative and the other positive. For our analysis the negative threshold corresponds to an "unacceptable" judgement and the positive threshold corresponds to an "acceptable" judgement. Thus, negative drift rates indicate movement toward the unacceptable response and the opposite holds for positive drift rates.

Looking at the mean parameter values depicted in Table 3, we observe that the *zr* parameter is incredibly close to 0.5 the halfway point between either threshold. The *a* parameter, threshold separation, is on high end of the expected range. This value is typically estimated to 2.0 or lower (Voss and Voss, 2007) which is exceeded here by a small margin. For v, the drift rate by condition, we see that the highest positive observed value is for the MultiMatch condition. This is followed by the smaller, but still positive LowMatch condition. Both the HighMatch and NoMatch conditions have negative average drift rates. The estimated value for the HighMatch condition is (absolutely) small relative to the other conditions and close to zero. The NoMatch condition has a stronge negative drift rate, one which is

close to the LowMatch condition in absolute magnitude.

To asses the effect of condition on drift rate, we conduct a 4x1 repeated measures ANOVA on the drift rates for each subject with condition as the factor. The sign of drift rates in the NoMatch condition are flipped across all tests for a more direct comparison of evidence quality. We observe a statistically significant effect ($F(3, 114) = 23.49282, p < 0.001$). To assess the effect of each condition, we perform six pairwise t-tests comparing drift rates for each possible combination of condition. We find significant differences ($p < 0.05$) for all conditions, except the NoMatch/LowMatch comparison ($t(38) = 0.96938, p = 0.3385$). This indicates that significance of the ANOVA test does not derive from any differences between the LowMatch and NoMatch conditions. Lastly, to asses the existence of response bias, we perform a one sample t-test using the estimated zr values across participants. A value of 0.5 for zr indicates the absence of response bias, thus we use $\mu = 0.5$ for this test. We observe no statistically significant difference ($t(38) = 0.39943, p = 0.6918$) between these means, indicating a lack of response bias in the estimated parameters.

6.2.4. Evaluation of Model Fit

We observe a mean of 0.634 and $p > 0.05$ for the p-values calculated by fast-dm. While this indicates a good quality fit, this method of is unreliable. Thus, as per the suggestion of Voss et al. (2013) we graphically represent model fit. In Figure 5 we plot the observed pattern of results against the pattern of results predicted by the model parameters estimated. We do this using a cumulative distribution function for all participants in each condition. The better match between the two functions, the better quality of fit achieved.

What we observe in Figure 5 is a strong overall qualitative match between the two functions. Across conditions the empirical and modeled CDF are close in both accuracy and distribution of reaction times. The match is especially strong for the HighMatch condition. One undesired pattern we observe is a consistent overestimation of judgement accuracy in the NoMatch, MultiMatch, and LowMatch conditions. Due to response latency judgements cannot occur at or near 0 seconds, thus no probability mass is accumulated within this region. Because this flat region lies between negative (unacceptable) values and positive (acceptable) values, a higher the percentage of data in this region indicates a higher percentage of unacceptable responses and vice-versa. Across the NoMatch, MultiMatch, and LowMatch conditions, we

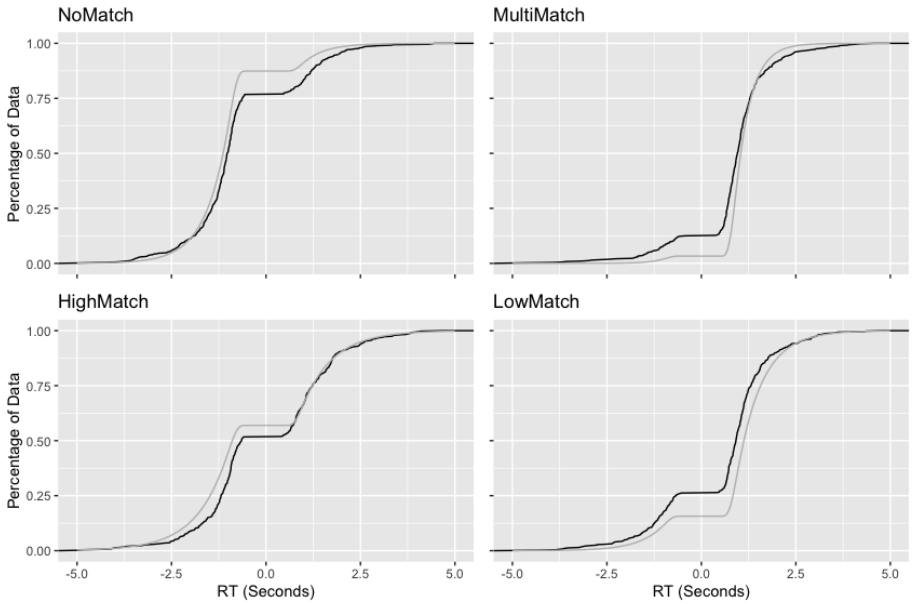

Figure 5: Empirical CDF (black) plotted against the Modeled CDF (grey) across all four conditions in Experiment 1; negative values indicate an "unacceptable" response, positive values indicate an "acceptable" response, and the absolute magnitude corresponds to reaction times in seconds

see that the modeled CDF predicts higher accuracy than is observed. The HighMatch condition is the only condition where this pattern does not hold.

Together the two measures utilized indicate a strong model fit. None of the p-values calculated by fast-dm reach significance, this is supported by a visual comparison of the empirical and modeled distribution of reaction times. The model does over predict response accuracy, but the difference is not tremendous and the qualitative pattern is captured across conditions.

6.3. Discussion

While some patterns persist, the results of experiment 1 fail replicate those of Dillon et al. (2019) in full.

The ambiguous MultiMatch condition is accepted faster, more often, and with less variation than both of the unambiguous conditions. This ambiguity advantage effect is expected under a parallel model where unambiguous sentences generate conflicting evidence at disambiguation. However, a parallel model also predicts strong performance for the NoMatch condition where

both parses are unacceptable. NoMatch accuracy is higher than either unambiguous conditions, but reaction times for correct NoMatch judgements are higher and more variable than the MultiMatch and LowMatch condition. At least under the parallel surprisal model, there is no reason an ambiguous ungrammatical sentence should exhibit this penalty. While these results are not not fully consistent with parallel processing, a serial model is not compatible either. If the parser has chosen the acceptable attachment in an unambiguous sentence, behavior should mirror the acceptable MultiMatch response. Instead, we observe that correct unambiguous responses are weaker than the MultiMatch condition across all measures. Given the time constraints enforced, it is unclear how a neglected parse could influence the perceived acceptability of a sentence under a serial model.

Barring acceptable responses to the ambiguous conditions, average reaction times and standard error have increased from Dillon et al. (2019). This is might be unsurprising given the change in task demands. Eliciting a judgement at disambiguation eliminates any window participants might have to determine acceptability. While this is reduces the possibility of rapid serial reanalysis, it might have the unintended consequence of shifting average reaction times higher. Accuracy for the grammatical conditions has risen along with reaction times. Given that the interpretability of our data hinges upon the feasibility of the task, this is an encouraging result. It seems that the time pressures enforced do not impede accuracy for grammatical sentences. Even in the NoMatch condition, where accuracy has fallen, the change is small and accuracy remains high. This lends confidence to the ambiguity advantage effect observed and the paradigm in general.

Performance for HighMatch is considerably weaker than other conditions, correct judgements are made close to chance with the slowest reaction times and the highest variation. Standard American English on average has a low attachment preference with respect to relative clauses, but this varies with respect to a number of factors (Gilboy et al., 1995). Dillon et al. (2019) norm their stimuli and find that participants prefer low attachments on average 70% of the time. This preference would impact the predictions of both parallel and serial models. Assuming equivalent grammaticality under the parallel suprisal model, a disprefered attachment will be less probable and yield higher surprisal than a preferred attachment. This predicts that performance will be worse for HighMatch than LowMatch because the grammatical parse is dispreferred in this condition. Serial models can select one parse from the available parses in a variety of ways, but a probable method might reflect

attachment preferences. This kind of serial model would select the preferred low attachment with greater frequency, making the LowMatch condition appear acceptable more often and the HighMatch condition less often. Both of these models fit the pattern observed. This explains the performance disparity between HighMatch and the other grammatical conditions, but why this difference has increased from Dillon et al. (2019) remains mysterious.

In drift diffusion, as with all modeling, the results are meaningless without first establishing the quality of fit. This a chief concern here as the application of the drift diffusion to speeded acceptability is novel- Hammerly et al. (2019) being the only prior work. Across all participants we observe non-significant p-values from the models estimated. This indicates a strong fit to the response data. In our graphical comparison of empirical and predicted reaction time distributions, we observe a strong qualitative match. Interestingly, accuracy is over estimated for some conditions, a pattern Hammerly et al. (2019) observe as well. The cause of this effect is unclear. However, as a whole these results inspire confidence in our use of the drift diffusion model and the parameter values we estimate.

Shifting to parameter results, the mean value estimated for zr is almost identical to the middle point. In our t-test we find no significant difference between a mean of 0.5 (the unbiased starting point) and the observed values. This is evidence that participants are not biased toward either choice in the judgement task. The mean a parameter is on the higher end of the typical range. This indicates that participants provide judgements conservatively, waiting for more evidence to accumulate. This might capture the increased average reaction times across conditions from Dillon et al. (2019). However, there is way to confirm this theory from the diffusion results alone.

In the drift rates we observe a mixed replication of Dillon et al. (2019) similar to the experimental data. While the highest absolute mean drift rates are found in the ambiguous conditions, the insignificance of the NoMatch/LowMatch t-test contradicts this pattern. A complete ambiguity advantage effect in line with parallel processing would involve significant differences between all ambiguous and unambiguous conditions. Further, under a parallel model we expect insignificant differences in drift within the ambiguous and unambiguous conditions. However, all of the t-tests aside from NoMatch/LowMatch reach significance. A serial model does not fit these results either. For unambiguous conditions evidence will always be strong, but it will pull toward a different thresholds depending on the parse chosen. For ambiguous conditions parses will always be consistent in acceptability, so

evidence will always pull strongly toward one threshold. While unambiguous drift rates are significantly different and reflective of attachment preferences, ambiguous drift rates differences are unexpectedly significant as well. Serial parsing does not predict the insignificance of the NoMatch/LowMatch comparison either.

In summary, we find a restricted ambiguity advantage effect and inconclusive evidence of parallel parsing strategies across experimental and modelling results. While we do observe an advantage for ambiguity, this is restricted to the grammatical conditions. There is no serial model we are aware of which can account for this pattern. Parallel models can capture this pattern, but offer no explanation for the weak NoMatch performance. It is unclear how the adjustments made from the Dillon et al. (2019) design might yield this penalty for ungrammatical sentences. The drift diffusion model achieves a strong fit, but the parameters are as conflicted as the data being modeled. We conclude that experiment 1 fails to replicate the findings of Dillon et al. (2019) and offers inconclusive evidence as to parallelism in parsing.

7. Experiment 2

The evidence for parallel parsing strategies has been conflicted across Dillon et al. (2019) and the first experiment. However, both of these experiments have dealt with parallelism in a broad sense and with good reason. The question these experiments intended to address is a fundamental one and abstraction from fine-grain details is necessary to answer such questions. With some evidence for and against a general parallelism, we now move on to address the predictions of a more developed parallel model.

The Gradient Symbolic Computation sentence processing model (Cho et al., 2017) is a parallel model which utilizes blended representations that gradually settle to a single parse over time. These blended representations are implemented with tenor products and capture basic phrase structure rules. The representations are a conjunctive blend of the syntactic representations which are consistent with the linguistic input received. There are two parsing principles of the model that warrant discussion: grammatical constraints and quantization. The grammatical quality of any point within the continuous space of a GSC representation is evaluated utilizing grammatical constraints within a harmonic grammar (Smolensky, 2006). A higher harmony indicates the satisfaction of more grammatical constraints. Quantization is the force which drives the model to settle into a discrete representation from a contin-

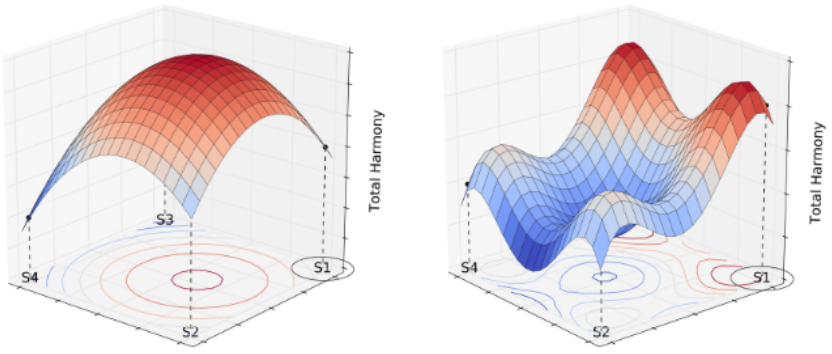

Figure 6: Harmony visualized for a continuous GSC representation

uous one. As time passes, the strength of quantization builds. This makes continuous GSC representations inherently unstable. The GSC parser arrives at a single discrete parse by means of stochastic gradient ascent. This can be described by visual analogy. Turning to Figure 6 we see two steps along the path of GSC parsing. On the left, we see the continuous representation at the introduction of structural ambiguity. There are four potential parses depicted in the four corners. We observe that initially, lacking both disambiguating material and a strong emphasis to quantize, the point of highest harmony is between all four parses. Gradient ascent will move toward this parallel-like peak. On the right, we see the continuous representation at a later state. Further evidence has weakened the harmony of some parses and strengthened the harmony of others. As well, quantization has strengthened the harmony of the four corners corresponding to a discrete parse. The S1 and S3 peaks are nearly equivalent harmonically, which is why gradient ascent occurs stochastically. In the case of a globally ambiguous sentence, parses will be close both in acceptability and harmony. The pressure to quantize will force one of the parses to be chosen probabilistically.

The GSC parser that Cho et al. (2017) implement utilizes a simple toy grammar to test these mechanics which have been outlined. The material which this model is currently capable of processing are much less complex than the relative clause ambiguities that have been our focus. However, the central principles of the GSC model can be applied to these constructions.

Gradient symbolic computation predicts that parallelism is inherently unstable and that time forces the parser to transition to more serial-like representations. As such, intervening material inserted between the introduction of a relative clause attachment ambiguity and disambiguation could generate more serial-like response patterns. We predict faster reaction times with less variation across conditions as representations become more serial- less uncertainty should yield less hesitation in responses. This should hold for both acceptable and unacceptable judgements. For the drift diffusion results this would predict similar drift rates across all conditions- less uncertainty should yield less variation in the quality of evidence. In the second experiment, we manipulate stimuli from the first experiment to test these predictions.

7.1. Method

7.1.1. Participants

Forty seven participants were recruited and utilized for the second experiment. All were recruited from an advertisement posted to Prolific. Participants were American, monolingual native English speakers, between the ages of eighteen and fifty, with no known language related disorders. Individuals who had participated in the first experiment were prohibited from participating in the second. Participants were compensated with $5.00 U.S. dollars. This procedure was reviewed and approved by the University of Massachusetts Amherst Internal Review Board.

7.1.2. Materials

All forty of the critical experimental items in experiment 2 were identical to those of the first experiment with one key manipulation- the addition of intervening material. Only four items from experiment 1 and Dillon et al. (2019) had one word between the complementizer and the disambiguating verb.

(5) Four conditions from experiment 2

a. NoMatch: Franny observed the nurses of the surgeons who, one time last year, was

b. MultiMatch: Franny observed the nurse of the surgeon who, one time last year, was

c. LowMatch: Franny observed the nurses of the surgeon who, one time last year, was

d. HighMatch: Franny observed the nurse of the surgeons who, one time last year, was

Intervening material was added to all forty critical items. This material was either a temporal modifier or a prepositional phrase, both in an adjunct position. This is depicted in Example 5. In either case, these modifiers were always between four to five words long and were always surrounded by commas. Length was kept consistent to ensure stability of any potential effect and the commas were utilized to provide a prosodic cue in writing. Additionally, all intervening material was marked singular so as to limit any potential agreement attraction effects at the disambiguating verb. Otherwise the critical items are identical. Similar intervening material was added to twenty of the sixty-eight fillers. These were balanced between grammatical and ungrammatical and all tested the same relative clause ambiguity utilized by the critical stimuli. This was done to prevent the intervening material from becoming a flag distinguishing the critical items from the filler items. The critical items were latin squared and combined with these fillers and each participant was exposed to an even split of grammatical and ungrammatical stimuli.

7.1.3. Procedure

The procedure was identical to that of the first experiment.

7.1.4. Fast-DM

We utilize fast-dm (Voss and Voss, 2007) version 30.2 (Voss et al., 2015) to calculate drift diffusion parameter estimations of our results. Again, we utilize the Kolmogorov-Smirnov (KS) optimization procedure for the balance between stability and efficiency of estimation.

7.2. Results

Of the 47 participants recruited 41 were utilized for analysis. One of these participants was rejected for having response accuracy on the filler items 2.5 standard deviations below the mean. The remaining 5 were rejected for having one or more judgement reaction times above 6 seconds for the critical items.

7.2.1. Accuracy

Judgement accuracy is closest among conditions which match in ambiguity. The ambiguous conditions are within 5% of one another and there is only a 1% difference in accuracy for the the unambiguous conditions. These values were much more distant in the first experiment. As in the first experiment, ambiguous conditions yield higher accuracy as compared to unambiguous conditions. Accuracy continues to be the highest for the MultiMatch condition, but by a smaller margin from the NoMatch condition in this experiment. LowMatch accuracy has decreased from the first experiment, while HighMatch accuracy has risen to meet at 60% in the middle. Accuracy as fallen for the ambiguous conditions as well, though more substantially for the MultiMatch condition.

7.2.2. Reaction Time

We observe, as before, the fastest and least variable reactions for correct MultiMatch responses. This is followed by correct LowMatch, HighMatch, and NoMatch responses. The differences in reaction time across correct responses are less drastic than in the first experiment. As a consequence of this, there is much more overlap across the condition when the standard error values are considered. Lastly, incorrect responses are provided slower on average and with variation across all conditions.

		"Acceptable"	**"Unacceptable"**
Condition	**Accuracy**	**Reaction Time (Standard Error)**	
NoMatch	72%	1338 (63)	1262 (37)
MultiMatch	77%	1109 (31)	1296 (67)
LowMatch	60%	1207 (43)	1400 (64)
HighMatch	59%	1256 (45)	1339 (57)

Table 4: Average reaction times in milliseconds and accuracy of judgements across all four conditions in Experiment 2, correct responses highlighted in grey

7.2.3. Drift Diffusion Parameter Estimations

Looking at parameter values, we see that zr is above the halfway point, indicating a bias toward the positive-coded acceptable response threshold. The a parameter remains on the higher end of the typical range (Voss and Voss, 2007), but has decreased from the first experiment. The qualitative pattern for the drift rate persists from the first experiment, MultiMatch is highest

Parameter	Mean	SE
a	1.875	0.069
zr	0.581	0.022
vNoMatch	-0.932	0.162
vMultiMatch	1.011	0.192
vLowMatch	0.322	0.206
vHighMatch	0.037	0.145

Table 5: Mean values and standard error of estimated drift diffusion parameters in experiment 2, drift rates are estimated by condition

followed by NoMatch, LowMatch, and lastly HighMatch. All grammatical conditions now have positive drift, as the HighMatch condition changed from slightly negative to slightly positive.

We utilize the same 4x1 repeated measures ANOVA on drift rate with condition as the factor. We continue to observe a statistically significant effect ($F(3, 120) = 6.948677, p < 0.001$), suggesting that our grammatical manipulations have a significant impact on drift rate. We perform the same six pairwise t-tests using the drift rates for each condition. We find significant effects ($p < 0.05$) for all comparisons except the LowMatch/HighMatch ($t(40) = -1.2546, p = 0.2169$) and MultiMatch/NoMatch ($t(40) = -0.32354, p = 0.748$) comparison. We take this results to indicate that conditions which are similar in ambiguity have similar drift rates. Lastly, we perform a one sample t-test with $\mu = 0.5$ and the zr values estimated to asses response bias. Unlike the first experiment, we find a significant difference ($t(40) = 3.6511, p < 0.001$) in our test. This indicates that there is a bias toward the acceptable response in experimental results.

7.2.4. Evaluation of Model Fit

We use the KS statistic derived p-value for each participant as a preliminary measure of model fit. We observe a mean p-value of 0.1693 and $p > 0.05$ for all participants- indicating that the model is a good fit to the results. To confirm this result, we graphically represent model fit as well. In Figure 7, we plot the empirical and modeled CDFs for all participants in each condition. We observe a strong resemblance between the two functions across all four conditions, the fit is particularly strong for the HighMatch condition. We continue to observe an overestimation of accuracy from Hammerly et al. (2019) and the first experiment. The HighMatch condition is the

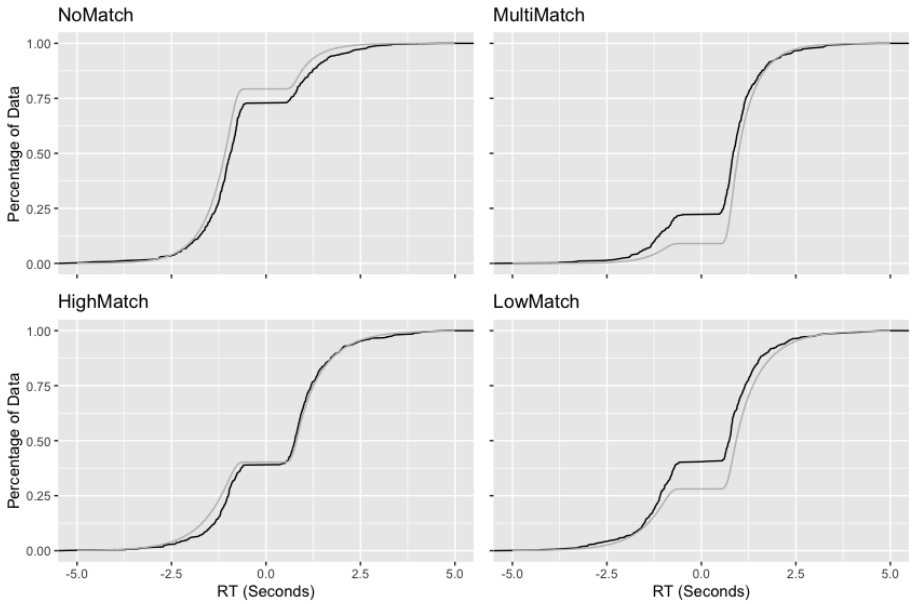

Figure 7: Empirical CDF (black) plotted against the Modeled CDF (grey) across all four conditions in Experiment 2; negative values indicate an "unacceptable" response, positive values indicate an "acceptable" response, and the absolute magnitude corresponds to reaction times in seconds

only condition where accuracy is not overestimated. On the basis of these two measures, the drift diffusion model continues to accurately describing behavioural results.

7.3. Discussion

By and large, the addition of intervening material does not appear to yield more serial behaviour as predicted by a GSC parser. Instead, what we observe in the second experiment is a response pattern comparable to that of the first experiment.

Reaction times and accuracy have changed dramatically. Across all conditions reaction times have fallen from the first experiment. Differences in reaction time, standard error, and accuracy across conditions are smaller as well. We might expect these results under a serial model where there is only one parse and correspondingly less uncertainty. However, counter to this, we continue to observe an ambiguity advantage effect. Accuracy is higher for ambiguous conditions than unambiguous conditions and by a larger margin than

in the first experiment. As in the first experiment, the ambiguity advantage is constrained to grammatical conditions. Correct MultiMatch responses are provided faster and with less variation than either correct LowMatch or HighMatch responses. Unlike the first experiment, correct NoMatch responses are provided slower than both unambiguous conditions. This grammatically restricted advantage effect is not predicted by a parallel, serial, or hybrid GSC model. In short, while there is an ambiguity advantage effect counter to a serial parser, weak performance in the NoMatch condition does not fit a parallel parser either.

Unexpectedly, LowMatch accuracy has fallen while HighMatch accuracy has risen- meeting at around sixty percent. One possible explanation is that the intervening material weakens the low attachment preferences of English speakers. There is no independent evidence for this specific effect, but English attachment preferences can be manipulated (Gilboy et al., 1995) and this material does add linear distance between the lower argument and the disambiguating verb. From the Dillon et al. (2019) norming results we know that a low bias is present in the original stimuli, but without norming our modified stimuli we cannot be certain if preferences have shifted. As discussed in the first experiment, attachment preferences could impact reaction times and accuracy in either a serial or parallel model. So even if preferences have shifted, this finding would not reveal parsing behaviour in the results.

Before analyzing the drift diffusion parameters, we evaluate the model fit to the response data. Looking at the cumulative distribution functions, we continue to observe an overestimation of judgement accuracy for the same three conditions as in the first experiment. Including Hammerly et al. (2019), there are now three speeded acceptability experiments which have found this overestimation. These experiments utilize the same paradigm, similar participant counts, and the same optimization criterion but it remains unclear how these shared factors might contribute to overestimation. Despite this we observe a strong qualitative match between the empirical and modeled density functions. The model is particularly accurate for the HighMatch condition. Lastly, while less reliable as a measure of fit, none of the p-values for the models we fit achieve significance. All of these results indicate that the drift diffusion model is able to account for the results of the second experiment. Given that model fit has been strong in both experiments, we are confident in the parameter values we estimate and in the application of the drift diffusion model.

The mean value for zr has shifted above the unbiased 0.5 from the first

experiment and our t-test yields a significant result. This indicates that participants are more biased to provide acceptable judgements. As such, we expect higher reaction times for conditions where drift is negative and participants provide unacceptable judgements. A pattern which we observe in the NoMatch condition. Threshold separation or a remains conservative, but falls from the first experiment. We estimate a mean value which is smaller, but still on the higher end of the typical range. This predicts that, as compared to the first experiment, participants will require less evidence to accumulate before providing a judgement. A prediction that is consistent with the fall in reaction times we observe across all conditions from the first experiment.

Absolute drift rates are lower for all conditions except HighMatch. In the first experiment, HighMatch drift was slightly negative on average and it is now slightly positive. These changes in drift seem to reflect the changes in accuracy, where LowMatch has increased the other conditions have fallen. Despite these shifts, we observe an ambiguity advantage within the absolute values of the mean drift rates. The highest drift rates are for the ambiguous NoMatch and MultiMatch conditions. The pairwise t-tests indicate that drift rates for the ambiguous conditions are significantly different from the unambiguous conditions. These tests also indicate that the differences within the ambiguous and unambiguous conditions are not significant. This is the exact pattern we would expect under a parallel model. When the two parses held in parallel align in acceptability, participants should receive similarly strong evidence for the judgement task regardless of the grammaticality of the sentence. When these two parses conflict in acceptability, participants should receive similarly weak evidence for the judgement task regardless of which parse for the sentence is acceptable.

While reaction times and standard error have fallen across all conditions, this is the only pattern whic is evocative of the serial processing that the GSC predicts. As well, this finding can be explained by the drift diffusion parameter values estimated. Threshold separation indicates that participants are less conservative which would predict lower reaction times. Response style can vary between participant samples and this seems more probable explanation than hybrid GSC representations. As in the first experiment, there is an ambiguity advantage effect constrained to grammatical conditions, a finding which is at odds with a parallel model. The diffusion modeling results offer a possible explanation. The relativized starting points estimated are found to be significantly different than an unbiased value. Thus, given equally

strong evidence, we expect higher reaction times to provide an unacceptable responses. Drift rates are not significantly different for the NoMatch and MultiMatch conditions, so we expect higher reaction times for correct NoMatch responses. HighMatch and LowMatch drift rates are not significantly different either. Given the significance of all other drift rate comparisons, the lack of significance with ambigous and unambiguous conditions makes a strong case for parallel parsing strategies. Considering the modeling and response data as a whole, we observe a better replication of Dillon et al. (2019) than the first experiment.

It is difficult to reconcile the differences we observe between the experimental and modelling results. On one hand, a key value of the drift diffusion model is its capacity to synthesize both reaction time and accuracy data to capture patterns which either measure alone might obfuscate (Voss et al., 2013). As such, the models we fit might match parallel processing more closely because that is the pattern underlying the experimental results. On the other hand, speeded acceptability is not a paradigm which the drift diffusion model has been widely utilized to capture. While we do observe indications of a strong model fit, we believe it would be rash to put the parameter estimations before the experimental results in our analysis. As such, we determine that our results are inconclusive.

8. Conclusion

Given our analysis of these experiments, we are faced with two options in explaining the relationship between them. One, the failure to fully replicate the results of Dillon et al. (2019) in the first experiment reflects serial processing and the addition of intervening material in the second second experiment yields more parallel processing. An issue with this account is explaining why delaying disambiguation would yield this effect. There is no theory which predicts a transition from serial to parallel parsing and there is no obvious reason that intervening material would yield this effect. Another problem with this account is that the results of the first experiment still match some predictions of a parallel model. While the high mean reaction time in the NoMatch condition is problematic, accuracy is still higher for this condition than either unambiguous condition. As well, we find an advantage for ambiguous material among the grammatical conditions- an effect which no serial model we are aware of can accommodate. As such, we find this account improbable. Two, the failure to fully replicate the results of Dillon

et al. (2019) in the first experiment reflects low power through which some traces of parallel processing persists. The addition of intervening material in the second experiment has no effect on processing strategies, but we do observe stronger evidence of parallel processing. A substantial amount of participants had to be removed from the analysis on account of high reaction times and the requirements of the drift diffusion model. In the first experiment we use results from 39 participants and in the second experiment we use results from 41 participants. This is on the lower end of what would be acceptable power for an acceptability experiment. As well, our design has a total of 40 trials per participant, this is on the lower end of the acceptable amount for drift diffusion modelling (Voss et al., 2013). Under these limitations of power some noise in the results might be expected. While this inspires less confidence in our results, it seems more probable than the first account which rejects patterns found across three separate experiments without an alternative explanation.

There were two main objectives for this research at two levels of interest. The first was to rerun the experiment from Dillon et al. (2019) with a series of adjustments aimed at yielding better controlled results. In doing so, we intended to replicate the ambiguity advantage effect and tease out a response pattern predicted by either serial or parallel models. What we find in the results of the first experiment is a mixed bag. Even when the possibility of reanalysis is eliminated and disambiguation and acceptability tasks are elicited in tandem, an ambiguity advantage persists. However, this advantage seems to be restricted to grammatical materials. This finding is troubling for both serial and parallel models and the drift diffusion results do not clarify these conflicted patterns. The second objective was to take the paradigm developed and utilize it to tackle predictions from a more developed parallel sentence processing model- the gradient symbolic computational parser. In this experiment, we find the opposite response pattern predicted under a theory of unstable parallel representations- a core aspect of the GSC model. Strangely with the addition of intervening material between complementizer and disambiguation, we observe a better replication of Dillon et al. (2019) and stronger evidence of parallel parsing strategies. Our evidence is in no way damming for the GSC model. There are a variety of reasons why GSC parsing strategies did not manifest. We may have made an incorrect choice of constructions, paradigm, length of interveening material, or linking function for example. In summary, we put forth a conflicted set of findings, some which support of parallelism and others which appear more serial in nature.

However, across none of our results do we observe patterns which match the predictions of a gradient symbolic computation parser.

9. Future Work

The most obvious way to follow up this research would be to run the same experiment with additional controls and participants. In the experiments Hammerly et al. (2019) run, when a participant takes longer than 3 seconds to provide a judgement the trial ends and they are instructed to provide judgements faster. Given that drift diffusion attempts explicitly to model the right tail of the reaction time distribution (Voss et al., 2013), we believe that is cut-off is too conservative. However, utilizing this same control with a more liberal 5 or 6 second limit seems wise. This would limit the number of trials that would need to be thrown out and fully ensure that participants are providing judgements in the way our analysis presupposes. Another beneficial change would be to have more trials per condition. Fast-dm requires a minimum of 10 trials per condition in order to estimate parameters by condition, with 40 trials and 4 conditions we meet this minimum exactly. As such, when we remove a trial which is above the reaction time limit, we must exclude an entire participant from analysis. Some amount of high reaction times are inevitable, especially with studies run online, thus writing more critical stimuli for each condition would be beneficial. Lastly, conducting a norming study for attachment preferences of the stimuli written for the second experiment is an obvious improvement. While these stimuli were only slightly modified from those utilized and normed in (Dillon et al., 2019), some of behavioural data in the second experiment might have emerged from a shift in attachment preferences. It is impossible to explore this possibility fully without norming these items.

Another avenue for future research would be to explore processing in this paradigm across languages and constructions. Although originally found in relative clauses, the ambiguity advantage effect has been demonstrated across a variety of constructions (Gompel et al., 2001; Grant et al., 2014; Gompel et al., 2005). Replicating this design with different constructions would offer a strong test of the paradigm and our results. As well, testing a diversity of constructions would benefit the evaluation of the GSC parser. It is impossible to say anything conclusive about this model having only tested one construction in a single experiment. Lastly, while English has a low attachment preference for relative clauses other languages have high attachment

preferences. While the significance of this effect varies, we continually observe weaker performance for the dispreferred high attachment condition in our experiment. It is unclear whether this results from the grammatical preferences of English speakers or the proximity of the relevant arguments to the disambiguating region. Running this same experiment in a high attachment language such as Spanish could prove informative.

References

B. Dillon, C. Andrews, C. M. Rotello, M. Wagers, A new argument for co-active parses during language comprehension, Journal of Experimental Psychology: Learning, Memory, and Cognition 45 (2019) 1271–1286. doi:10.1037/xlm0000649.

R. Ratcliff, A theory of memory retrieval, Psychological Review 85 (1978) 59–108. doi:10.1037/0033-295x.85.2.59.

P. W. Cho, M. Goldrick, P. Smolensky, Incremental parsing in a continuous dynamical system: sentence processing in gradient symbolic computation, Linguistics Vanguard 3 (2017) 1. doi:10.1515/lingvan-2016-0105.

M. J. Traxler, M. J. Pickering, C. Clifton, Adjunct attachment is not a form of lexical ambiguity resolution, Journal of Memory and Language 39 (1998) 558 – 592. URL: http://www.sciencedirect.com/science/article/pii/S0749596X98926006. doi:https://doi.org/10.1006/jmla.1998.2600.

L. Frazier, On comprehending sentences: Syntactic parsing strategies, Ph.D. thesis, Indiana University Linguistics Club, 1979.

L. Frazier, Sentence processing: A tutorial review, in: M. Coltheart (Ed.), Attention and Performance XII, Erlbaum, Hillsdale, NJ, 1987, pp. 559–586.

M. C. MacDonald, Probabilistic constraints and syntactic ambiguity resolution, Language and Cognitive Processes 9 (1994) 157–201.

K. McRae, M. J. Spivey-Knowlton, M. K. Tanenhaus, Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension, Journal of Memory and Language 38 (1998) 283–312.

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

E. Gibson, N. J. Pearlmutter, Distinguishing serial and parallel parsing, Journal of Psycholinguistic Research 29 (2000) 231–240. doi:10.1023/A:1005153330168.

W. Tabor, S. Hutchins, Evidence for self-organized sentence processing: Digging-in effects., Journal of Experimental Psychology: Learning, Memory, and Cognition 30 (2004) 431–450. doi:10.1037/0278-7393.30.2.431.

C. Clifton, L. Frazier, Comprehension of sluiced sentences, Language and Cognitive Processes 13 (1998) 499–520. doi:10.1080/016909698386474.

F. Ferreira, J. M. Henderson, Recovery from misanalyses of garden-path sentences, PsycEXTRA Dataset (1991). doi:10.1037/e665402011-369.

B. Dillon, A. Staub, J. Levy, C. Clifton Jr, Which noun phrases is the verb supposed to agree with?: Object agreement in American english, Language 93 (2017) 65–96. doi:10.1037/xlm0000649.

C. Hammerly, A. Staub, B. Dillon, The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence, Cognitive Psychology 110 (2019) 70–104. doi:10.1016/j.cogpsych.2019.01.001.

A. Voss, J. Voss, Fast-dm: A free program for efficient diffusion model analysis, Behavior Research Methods 39 (2007) 767–775. doi:10.3758/bf03192967.

A. Voss, J. Voss, V. Lerche, Assessing cognitive processes with diffusion model analyses: a tutorial based on fast-dm-30, Frontiers in Psychology 6 (2015). doi:10.3389/fpsyg.2015.00336.

A. Voss, M. Nagler, V. Lerche, Diffusion models in experimental psychology, Experimental Psychology 60 (2013) 385–402. doi:10.1027/1618-3169/a000218.

A. Voss, K. Rothermund, J. Voss, Interpreting the parameters of the diffusion model: An empirical validation, Memory & Cognition 32 (2004) 1206–1220. doi:10.3758/bf03196893.

A. Kolmogoroff, Confidence limits for an unknown distribution function, The Annals of Mathematical Statistics 12 (1941) 461–463. doi:10.1214/aoms/1177731684.

E. Gilboy, J.-M. Sopena, C. Clifton, L. Frazier, Argument structure and association preferences in spanish and english complex nps, Cognition 54 (1995) 131 – 167. URL: http://www.sciencedirect.com/science/article/pii/001002779400636Y. doi:https://doi.org/10.1016/0010-0277(94)00636-Y.

P. Smolensky, Harmony in linguistic cognition, Cognitive Science 30 (2006) 779–801. doi:10.1207/s15516709cog0000$_7$8.

R. P. V. Gompel, M. J. Pickering, M. J. Traxler, Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models, Journal of Memory and Language 45 (2001) 225–258. doi:10.1006/jmla.2001.2773.

M. Grant, B. Dillon, S. Sloggett, Ambiguity advantages in attachment and pronominal reference: Evidence from eye movements during reading, Paper presented at the 20th annual Architectures and Mechanisms for Language Processing Conference, University of Edinburgh, United Kingdom. (2014).

R. P. V. Gompel, M. J. Pickering, J. Pearson, S. P. Liversedge, Evidence against competition during syntactic ambiguity resolution, Journal of Memory and Language 52 (2005) 284–307. doi:10.1016/j.jml.2004.11.003.