A Multinomial Processing Tree Model of RC Attachment

Pavel Logačev
Boğaziçi University
Department of Linguistics
34342 Istanbul, Turkey
`pavel.logacev@gmail.com`

Noyan Dokudan
Boğaziçi University
Department of Linguistics
34342 Istanbul, Turkey
`noyan.dokudan@gmail.com`

Abstract

In the field of sentence processing, speakers' preferred interpretation of ambiguous sentences is often determined using a variant of a discrete choice task, in which participants are asked to indicate their preferred meaning of an ambiguous sentence. We discuss participants' degree of attentiveness as a potential source of bias and variability in such tasks. We show that it may distort the estimates of the preference of a particular interpretation obtained in such experiments and may thus complicate the interpretation of the results as well as the comparison of the results of several experiments. We propose an analysis method based on multinomial processing tree models (Batchelder and Riefer, 1999) which can correct for this bias and allows for a separation of parameters of theoretical importance from nuisance parameters. We test two variants of the MPT-based model on experimental data from English and Turkish and demonstrate that our method can provide deeper insight into the processes underlying participants' answering behavior and their interpretation preferences than an analysis based on raw percentages.

1 Introduction

One of the key questions in the field of sentence processing has been: *What does the human sentence processing mechanism do when confronted with an ambiguity?* A variety of different proposals regarding online disambiguation strategies have been made over the years, such as the Garden-path Theory (Frazier, 1987), the Tuning Hypothesis (Cuetos et al., 1996), the Competition-Integration Model (McRae et al., 1998) and many others. Their diverging predictions have led to a significant body of empirical research documenting, among other things, substantial cross-linguistic variation in the interpretation of ambiguous sentences: For instance, Cuetos and Mitchell (1988) compared the RC attachment preferences of English and Spanish speakers in ambiguous sentences like (1) and (2), in which the relative clause *'who had an accident'* can attach either to the NP headed by the first noun (N1, *'daughter'*) or to the NP headed by the second noun (N2, *'colonel'*).[1]

Cuetos and Mitchell presented Spanish-speaking and English-speaking participants with ambiguous sentences like (1) and (2) and asked them to answer comprehension questions like *'Who had an accident?'*. Participants' responses indicated that English sentences like (1) were assigned an N2 interpretation in 61% of the cases, while their Spanish counterparts like (2) were assigned an N1 interpretation in 72% of the cases. The authors interpret this finding as an argument against a cross-linguistically universal parsing strategy in the resolution of RC attachment ambiguities.

(1)　　The journalist interviewed the daughter$_{N1}$ of the colonel$_{N2}$ [who had an accident].

(2)　　El　periodista entrevisto　a la　hija$_{N1}$
　　　 The journalist　interviewed to the daughter
　　　 del　　coronel$_{N2}$ [que tuvo el accidente].
　　　 of the colonel　　[who had an accident].

Although disambiguation strategies seem to be at least partially determined by the linguistic properties of a given language, various other factors appear to influence the resolution of RC attachment ambiguities. For example, in a questionnaire study, Gilboy et al. (1995, inter alia) demonstrated a substantial influence of construction type. They asked participants to indicate which of the two available noun phrases was modified by the RC in several constructions. They found that the percentage of N2 attachment responses ranged between approximately 20% to 70% for their English sentences and

---

[1]To any ambiguity in the context of typologically diverse languages, we will refer to the two interpretation options as *N1 attachment* and *N2 attachment*, with N1 and N2 referring to the order of occurrence of the noun phrases head nouns instead of the more common terms *high attachment* and *low attachment*.

between 10% to 80% for their Spanish sentences. Grillo et al. (2015) also conducted a two-alternative forced-choice (2AFC) task in which English speakers choose between N1 and N2 as the attachment sites for the RC to indicate their interpretation of the sentence. They showed that English speakers, who had previously been claimed to prefer N2 attachment, preferred N1 attachment in more than 50% of the cases when a small clause reading was possible.

RC attachment preferences have also been studied in Turkish, where the order of the RC and the complex noun phrase is reversed, compared to English and Spanish. In a questionnaire study with sentences like (3), Kırkıcı (2004) found that animacy may affect attachment preferences such that when both NPs were [+human], there was no significant difference between the proportions of the N1 and N2 attachment, while an N1 attachment manifested when both NPs were [-human]. Contrary to this finding, Dinçtopal-Deniz (2010) found an across-the-board preference for N1 attachment in Turkish. In her questionnaire study, monolingual Turkish speakers read Turkish sentences with ambiguous RC-attachment and answered questions about them by indicating one of two options on each trial. The results of this study showed that participants preferred N1 attachment over N2 attachment: 66% percent of the responses indicated an N1 interpretation of the sentence.

(3) Şoför [şehir merkezin-de oturan]$_{RC}$
driver city   center-in   living
profesörün$_{N1}$ sekreterini$_{N2}$ gördü.
professor's   secretary   saw
'The driver saw the secretary of the professor who was living in the city center.'

2 The Role of Guessing

What most of the above studies of RC attachment preferences have in common is that they use some variant of a discrete choice task, in which participants select one of two response options to indicate their interpretation of the ambiguity. The relative proportion of responses indicating N1 and N2 attachment, respectively are interpreted as estimates of the magnitude of N1 or N2 attachment. A potential complication in interpreting the percentage of responses favoring an alternative in this way is that participants' responses may not always reflect their interpretation. At least on some trials, participants may process the sentence only partially or

fail to pay attention to it altogether. In such cases, participants' question responses must be based on an incomplete or nonexistent representation, and are more likely to resemble guesses than informed responses.

Evidence for such incomplete processing comes from the widely known fact that participants' accuracy in experimental tasks is often far from perfect, even for relatively simple tasks such as acceptability judgments: For example, Dillon and Wagers (2019) found in an *offline* acceptability judgment study that ungrammatical sentences like (4) are judged acceptable on 18% of the trials. Since it appears unlikely that sentences like (4) are considered grammatical and interpretable when fully processed, the explanation for such responses must lie in their incomplete processing followed by guessing.

(4) *Who do you think that the new professor is going to persuade anyone?

One way of conceptualizing a simple generative model of erroneous responses in relatively simple discrete choice tasks is to assume that at least some participants on some occasions fail to pay attention to the stimulus, and as a result, select a random response. If so, the relation between the probability of response X being actually preferred to alternative responses (p_X) and the probability of observing response X (p'_X) can be formalized as in equation 1: p'_X is the weighted average of (i) the probability of X being preferred to the alternative when the stimulus is fully attended to (p_X) and (ii) the probability of selecting X when the stimulus is *not* attended to (g_X), where a is the probability of attending to the stimulus.

$$p'_X = a \cdot p_X + (1-a) \cdot g_X \qquad (1)$$

Equation 1 illustrates that under the above assumptions, the proportion of responses indicating a preference for X conflates multiple factors. As a result, many preference estimates for X (p'_X) are compatible with a wide range of underlying preferences (p_X) under different assumptions regarding participants' degree of attentiveness and guessing behavior (a and g_X).

Table 1 illustrates this problem. It shows several parameter combinations which can account for a preference of 65% for X in a binary choice task. Such a finding may reflect (i) the absence of an

	p_X	a	g_X	p'_X
2	0.5	0.7	1	0.65
1	0.9	0.7	0.06	0.65
3	0.1	0.35	0.945	0.65

Table 1: Example combinations of parameters that may lead to an observed preference of approximately 65% according to equation 1.

underlying preference (table 1, row 1), (ii) the presence of a much stronger preference (table 1, row 2), and (iii) even a strong preference towards the alternative to *X* (table 1, row 3).

Given that participants in most if not all psycholinguistic tasks produce a sizeable amount of erroneous responses, it appears *a priori* quite plausible that such mechanisms are also at play in attachment preference studies. This means that empirical estimates of attachment preferences (p'_X) are likely to be (i) *biased towards the guessing parameter* g_X to a degree determined by a, and (ii) are likely to *vary between studies* as a function of the between-study differences in a and g_X. In the following, we propose a method for disentangling the contributions of attachment preferences and guessing using multinomial processing tree models (MPT; Erdfelder et al., 2009; Batchelder and Riefer, 1999) based on response patterns in unambiguous baseline sentences. We will first assess the empirical adequacy of two alternative MPT models on two experiments in English and Turkish, in which participants answered polar comprehension questions about sentences with ambiguous and unambiguous RC attachment. We will then compare the two experiments with regard to the parameter estimates obtained from the MPT models.

3 Experiments

To evaluate our method, which will be presented in the next section, we used question-answering data from two experiments in which participants read sentences with ambiguous and unambiguous RC attachments and answered polar comprehension questions about them.

3.1 Experiment 1

We used the RC question-answering data from Swets et al.'s (2008) self-paced reading experiment in English (N=48). In this experiment, participants read sentences like (5) in three attachment conditions and answered comprehension questions about

RC attachment similar to (6) on every trial. All comprehension questions required a *'yes'/'no'* answer. One-half of the questions asked whether the RC modified the noun phrase headed by N1, and the other half asked about N2.

RC attachment was disambiguated by means of gender (mis)match between the reflexive in the RC and the RC head noun. Each participant read 36 experimental sentences. Unambiguous sentences had correct answers, while the responses to ambiguous sentences indicated how readers disambiguated the sentence, thus reflecting their RC attachment preference.

(5)   a. AMBIGUOUS ATTACHMENT
The maid$_{N1}$ of the princess$_{N2}$ [who scratched *herself* in public] . . .

       b. N1 ATTACHMENT
The son$_{N1}$ of the princess$_{N2}$ [who scratched *himself* in public] . . .

       c. N2 ATTACHMENT
The son$_{N1}$ of the princess$_{N2}$ [who scratched *herself* in public] . . .

. . . was terribly humiliated.

(6)   COMPREHENSION QUESTION
Did the maid/princess/son scratch in public?

Figure 1 (left panel) shows the average percentages of *'yes'* responses to comprehension questions by attachment condition and question type (questions about N1 or N2).

3.2 Experiment 2

The second set of question-answering data came from an unpublished self-paced reading experiment on RC attachment in Turkish (N=99). In an experimental design similar to Swets et al., participants read sentences like (7). Because Turkish relative clauses are pre-nominal, the RC *who hit each other* preceded the complex noun phrase *the fans of the football players*. All RCs contained a reciprocal anaphor (*each other*), which allowed us to disambiguate the RC attachment by means of number marking on the head nouns as RCs with the reciprocals can only modify plural noun phrases. When only one of the nouns was plural, the sentence was unambiguous, and ambiguous when both nouns were plural since they were both licit attachment sites for the RC.

Participants were asked *'yes'/'no'* comprehension questions, like (8), which were always about

RC attachment. The comprehension question asked about the event mentioned in the RC and whether one of the nouns was involved in that event. Each participant read 42 experimental sentences. One-half of the questions asked whether the RC modified the noun phrase headed by N1, and the other half asked about N2. The experiment was conducted online on *ibexfarm* (Drummond, 2013). All participants were undergraduate students at Boğaz-içi University and native speakers of Turkish. Figure 1 (right panel) shows the average percentages of *'yes'* responses to comprehension questions by attachment condition and question type (question about N1 or N2).

(7) Dün   akşam, [birbirini   döven]$_{RC}$ ...
    Yesterday evening, each other hit

   a. AMBIGUOUS ATTACHMENT

     futbolcu-lar-ın$_{N1}$ hayran-lar-ı$_{N2}$ ...
     footballer-PL-GEN fan-PL-POSS

   b. N1 ATTACHMENT

     futbolcu-lar-ın   hayran-ı   ...
     footballer-PL-GEN fan.SG-POSS

   c. N2 ATTACHMENT

     futbolcu-nun   hayran-lar-ı ...
     footballer.SG-GEN fan-PL-POSS

     ... stadyumu hemen   terk etti.
     stadium   immediately leave did.

   'The fan(s) of the football player(s) who hit each other left the stadium immediately, yesterday evening.'

(8) COMPREHENSION QUESTION
    Futbolcu(lar)/hayran(lar)   dövüşte   yer almış mı?
    'Was/were the football player(s)/fan(s) involved in the fight?'

3.3 Results

The average percentages of *'yes'* responses in figure 1 indicate a substantial number of errors in unambiguous experimental conditions in both experiments, such as *'no'*-responses to N1 questions and *'yes'*-responses to N2 questions about N1 attachment sentences.

The average accuracy in answering questions about unambiguous sentences was 79% ($SE = 1.3$) in Swets et al.'s English experiment, and 66.5% ($SE = 2.5$) in our Turkish experiment.

The responses in the ambiguous attachment conditions indicate an N2 attachment preference in the English as 58% ($SE = 2.1$) of the response were
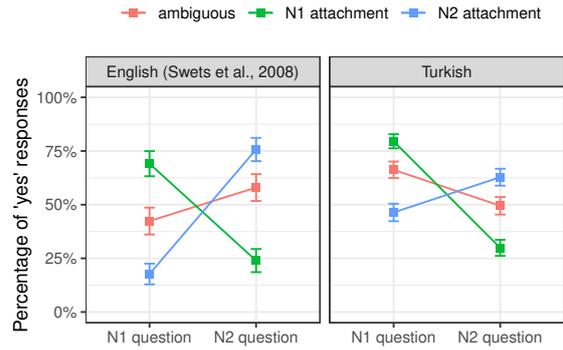

Figure 1: Average percentages of *'yes'* responses by attachment condition (color) and question type (x-axis). Error bars indicate 95% within-subject CIs.

compatible with N2 attachment (*'yes'* responses to N2 questions and *'no'* responses to N1 questions). Meanwhile, the Turkish data indicated an N1 preference as 58% ($SE = 1.9$) of the question responses were compatible with an N1 interpretation of the sentence. In both cases, the preferred attachment option is *local*, i.e., adjacent to the RC and is consistent with prior research.

Even though the estimates of the magnitude of the attachment preference are coincidentally equal, the magnitude of the preference for local attachment may not be. This is due to the presence of a substantial number of erroneous responses in unambiguous conditions in both experiments. Their presence indicates a substantial number of *guessing trials*, and thus suggests that not all N1- or N2-compatible responses in ambiguous indicate that the participant has successfully formed an N1- or N2 attachment interpretation of the sentence as they may have been generated by the same extraneous cognitive process that generates erroneous responses in the unambiguous attachment conditions.

The problem is exacerbated by the fact the response accuracy is particularly low in the N2 attachment condition in Experiment 2 (58.2%). A possible reason for this is that even on trials resulting in an N2 interpretation, the parser always attempts to construct an N1 attachment structure first because, in Turkish, unlike in English, potential attachment sites are processed sequentially after the relative clause has already been processed. As a result, the presence of a discarded alternative N1 attachment structure (e.g., Staub, 2007) could interfere with the retrieval of the correct structure during question answering in N2 attachment conditions. If,

as a result of retrieval failure, participants resort to guessing, we would expect to observe a substantial number of erroneous responses following N2 attachment sentences or ambiguous sentences which were ultimately disambiguated towards N2 attachment.

In the next section, we present two models of erroneous responses and then use them to estimate the magnitude of the actual strength of the attachment preference.

4 MPT Models of Question-Answering and Attachment

In accounting for the influence of extraneous cognitive processes, we considered two mechanisms that may generate erroneous question responses, and implemented both as multinomial processing tree (MPT) models (Batchelder and Riefer, 1999). In the following sections, we will use the model with the better empirical fit to obtain less biased estimates of the attachment preferences in the ambiguous conditions.

MPT models offer a way to formalize hypotheses about how a mixture of several latent processes generates a categorical response (cf. Erdfelder et al., 2009, for an overview). That is, under the assumption that different sequences of events may occur on different trials, the latent processes hypothesized to be involved in processing are represented as a probability tree, with each path through the processing tree corresponding to unique combinations of cognitive processes which give rise to a particular response, along with the probabilities of each path. Importantly, this formalization provides a framework in which the probabilities of relevant latent processes can be estimated. We will use them to estimate the magnitude of the RC attachment preference in Turkish and English.

4.1 Model 1

The first mechanism we considered as a potential explanation for erroneous responses is that participants sometimes fail to attend to or successfully process the stimulus or the comprehension question and simply press a random button. We hypothesize that this may happen due to inattentiveness, careless responding, distractions in the environment, mind-wandering (e.g., Smallwood, 2011), (temporary) fatigue, or failure to allocate sufficient processing resources towards the experimental task. We will subsume all of these factor under the um-

brella term *inattentiveness*.

The failure to process the stimulus is assumed to affect all three attachment conditions to the same degree. When participants do successfully comply with the task, they always respond to comprehension questions correctly in unambiguous conditions, while in ambiguous conditions, they sometimes adopt an N1 attachment interpretation of the sentence, and sometimes an N2 attachment, and answer comprehension questions in accordance with the adopted disambiguation of the ambiguous structure.
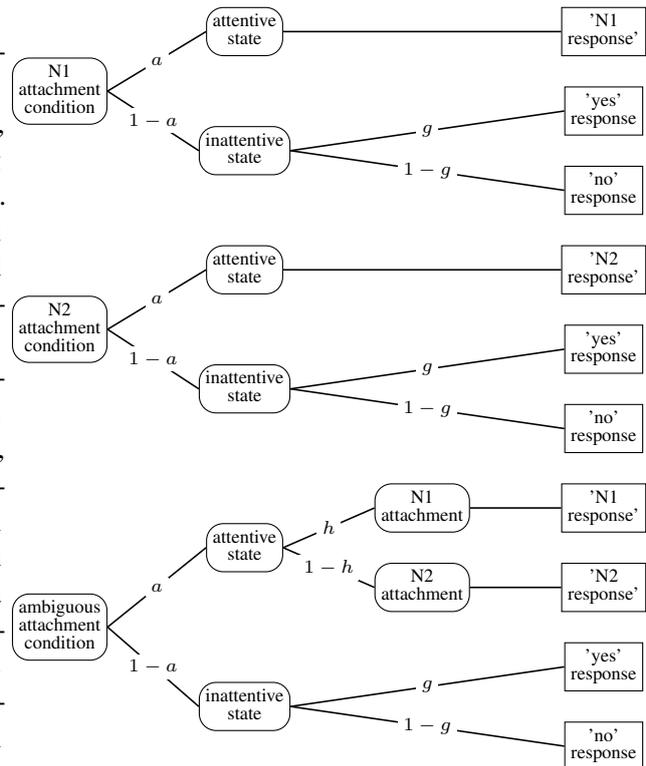
Figure 2: An MPT model of question answering with equal error rates for (i) N1 attachment, (ii) N2 attachment, and (iii) ambiguous sentences.

The assumptions of this account are illustrated in figure 2. The processing tree at the top illustrates how events during the processing of an N1 attachment sentence can unfold: On any given N1 attachment trial, a participant may be in an attentive state (with probability a) or an inattentive state (with probability $1 - a$). If the participant is in an attentive state throughout the trial (i.e., during reading and question answering), they will form a memory trace of the sentence they read, and later use it to correctly answer a comprehension ques-

43

tion. This is illustrated in the top branch of the N1 attachment condition schematic in figure 2, where *'N1 response'* stands for *'yes'* responses to N1 questions and *'no'* responses to N2 questions.

If the participant is in an inattentive state at any point during the trial (i.e., during reading or question answering), they will either fail to form a memory trace of the sentence they read or will fail to use it to answer the comprehension question. On those occasions, they will respond *'yes'* with probability g, and *'no'* with probability $1 - g$. This is illustrated in the bottom branch of the N1 attachment condition MPT schematic in figure 2.

As a result of these assumptions, the probability of a *'yes'* response in the N1 attachment condition is as given in equation 3, where I_{N1} (as in eq. 2) is an indicator variable which is 1 for N1 comprehension questions (such as *'Did N1 do RC?'*) and 0 for N2 comprehension questions such (as *'Did N2 do RC?'*).

$$I_{N1} = \begin{cases} 1, & \text{for trials with N1 questions} \\ 0, & \text{for trials with N2 questions} \end{cases} \quad (2)$$

$$p_{Y|N1} = a \cdot I_{N1} + (1 - a) \cdot g \quad (3)$$

The processing assumptions for the N2 attachment (middle, figure 2) condition and the ambiguous condition (bottom, figure 2) follow a similar logic, with the probability of a *'yes'* response given by equations 4 and 5.

$$p_{Y|N2} = a \cdot (1 - I_{N1}) + (1 - a) \cdot g \quad (4)$$

An important assumption about the hypothesized processes in ambiguous attachment conditions is that when readers are in an attentive state, they disambiguate ambiguous sentences either towards an N1 interpretation (with probability h) or an N2 interpretation (with probability $1 - h$). We make no assumptions about whether that happens during reading or at the question answering stage.

$$p_{Y|A} = a \cdot [h \cdot I_{N1} + (1-h) \cdot (1-I_{N1})] + (1-a) \cdot g \quad (5)$$

Importantly, we make no assumptions as to what may bring on inattentiveness and whether it occurs predominantly during reading or question answering. The key assumption of this account, however, is that this process affects all attachment conditions to the same degree.

4.2 Model 2

The second model included an additional possible source of erroneous responses that may not affect all attachment conditions equally. We hypothesized that, as observed in the unambiguous conditions of Experiment 2, one of the two interpretations (N1 or N2 attachment) could be more prone to failure, in that it may be less likely to be successfully created during reading, or less likely to be successfully recalled during question answering.
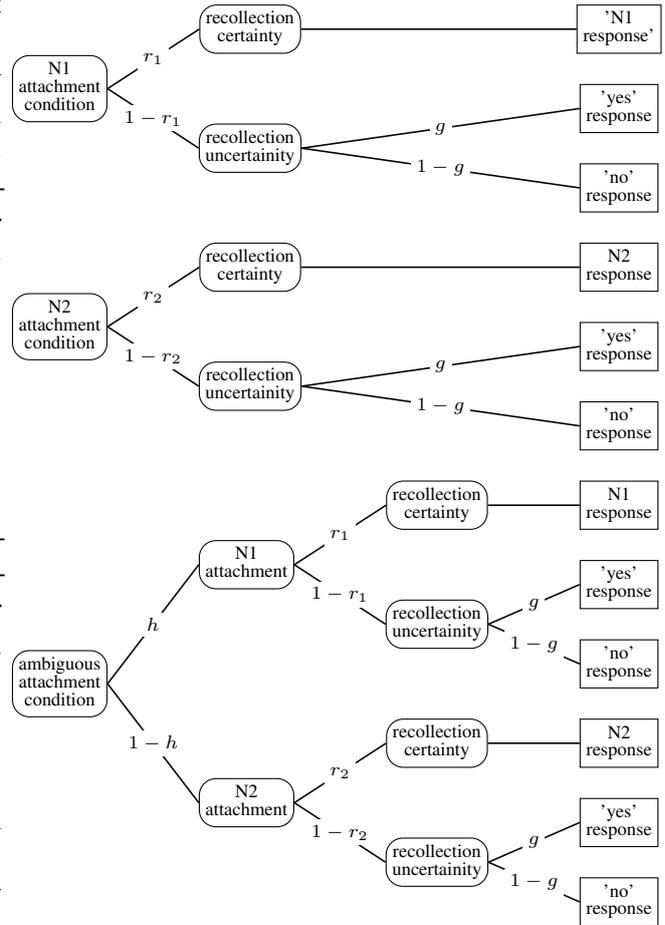

Figure 3: An MPT model of question answering with different error rates for N1 attachment and N2 attachment processes.

We formalized the assumption of different error rates associated with N1 and N2 attachment in the model in figure 3. The hypothesized structure of unambiguous N1 and N2 attachment trials is similar to Model 1 in figure 2. Each attachment process (N1 and N2 attachment) is associated with a probability of complete recollection certainty (r_1 and r_2, respectively) which reflects the probability that the

correct sentence structure is (i) constructed during reading and (ii) later correctly recalled during the question answering phase. If the correct sentence structure is constructed and recalled, participants respond in accordance with the structure they constructed. Otherwise, they select a random response, i.e., *'yes'* with a probability of g and *'no'* with a probability of $1 - g$. The probability of a *'yes'* response for all attachment conditions is given in equations 6, 7, 8.

In the ambiguous condition (figure 3, bottom), the recollection certainty and recollection uncertainty nodes are nested under the RC attachment nodes because the probabilities of the recollection certainty and uncertainty states depend on which RC attachment was chosen.

$$p_{Y|N1} = r_1 \cdot I_{N1} + (1 - r_1) \cdot g \qquad (6)$$

$$p_{Y|N2} = r_2 \cdot (1 - I_{N1}) + (1 - r_2) \cdot g \quad (7)$$

$$p_{Y|A} = h \cdot p_{Y|N1} + (1 - h) \cdot p_{Y|N2} \qquad (8)$$

Importantly, Model 2 (fig. 3) subsumes Model 1 (fig. 2) and therefore does not exclude the influence of an additional attention-related processes that affect all attachment conditions equally. This is because it can be re-parameterized as $(1 - r_1) = (1 - a) + (1 - r_1')$ and $(1 - r_2) = (1 - a) + (1 - r_2')$, such that the guessing rates in N1 and N2 attachment conditions, $(1 - r_1)$ and $(1 - r_2)$, can be interpreted as the sums of the attention-related guessing rate $(1 - a)$ and the condition-specific guessing rates $(1 - r_1')$ and $(1 - r_2')$.

5 Method

We implemented both MPT models[2] in *brms* and *rstan* (Bürkner, 2018; Stan Development Team, 2020) in R (R Core Team, 2018) according to equations 3-8. We fitted the models to each experiment separately, using 4 MCMC chains with $1,000$ warm-up and $3,000$ post-warm-up iterations. For the sake of computational convenience, we estimated all model parameters on the logit scale, and in the following, we will use θ' to refer to the logit-transform of any parameter θ.

We used mildly informative Gaussian priors for all logit-transformed population parameters in both models: $h', g' \sim N(0, 1)$, and $a, r_1, r_2 \sim N(0, 1)$.
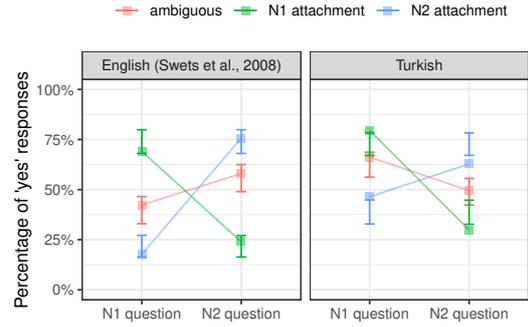
Figure 4: Average percentages of *'yes'* responses in the experiments, and 95% posterior prediction intervals based on Model 1 by attachment condition and question type.
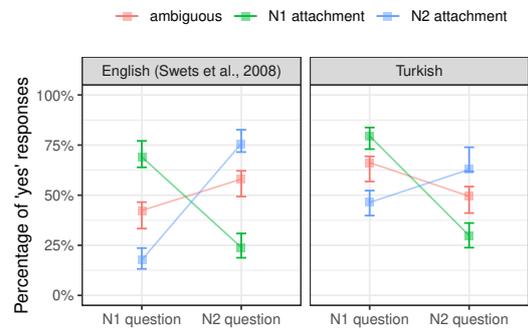

Figure 5: Average percentages of *'yes'* responses in the experiments, and 95% posterior prediction intervals based on Model 2 by attachment condition and question type.

To account for individual differences in all parameters, we used hierarchical models with by-subject intercepts for all parameters, where each participant k's responses were modeled as a function of population-level parameters θ with subject subject-level adjustments $\delta_{\theta,k}$, with $\theta_k' = \theta' + \delta_{\theta',k}$, where the by-subject adjustments are distributed as $\delta_{\theta',k} \sim N(0, \sigma_{\theta'})$.

6 Model Comparison

Figures 4 and 5 show the average percentages of *'yes'* responses by experiment (circles and connecting lines) alongside 95% posterior predictive intervals generated by, Model 1 and 2, respectively (error bars).

Figure 4 shows that although Model 1 could approximate the experimental findings it systematically overestimated the proportion of responses compatible with the preferred RC attachment (N2 in English, N1 in Turkish) in both unambiguous conditions: For example, in the N1 attachment con-

	English	Turkish
	$\widehat{elpd}$	$\widehat{elpd}$
model 1	$-511.3\ (18.1)$	$-796.4\ (15.0)$
model 2	$-469.4\ (14.8)$	$-750.5\ (13.5)$
	$\widehat{\Delta elpd}$	$\widehat{\Delta elpd}$
model 2-1	$41.9\ (11.6)$	$45.9\ (9.3)$

Table 2: Estimates of expected log pointwise predictive density ($\widehat{elpd}$) by model for each experiment and differences between model $\widehat{elpd}$s. Standard errors in brackets.

Figure 6: Population parameter estimates and 95% credible intervals for all four parameters of Model 2 (g, h, r_1, r_2) for both experiments, English (EN) and Turkish (TR).

dition in English, the number of *'yes'* responses to N1 questions and *'no'* responses to N2 questions were slightly overestimated. Similarly, in the N2 attachment condition in Turkish, the model underestimated the percentages of *'yes'* responses to N1 questions and *'no'* responses to N2 questions. Figure 5 shows that Model 2 appeared to have fewer systematic deviations, and appeared to fit the data quite well.

In order to compare the models more formally, we using PSIS-LOO-CV (Vehtari et al., 2017) to compute each model's expected log pointwise predictive density (ELPD). ELPD provides an estimate of the model's out-of-sample performance and thus penalizes additional model flexibility, which puts Models 1 and 2 on an equal footing although Model 2 has more parameters. Table 2 shows the ELPD estimates ($\widehat{\Delta elpd}$), as well as the differences between models in $\widehat{\Delta elpd}$ along with their respective standard errors. Larger values indicate better performance.

Both $\widehat{\Delta elpd}$ estimates are relatively large relative to their standard errors, and thus point towards Model 2 having better out-of-sample performance. This finding suggests that the two attachment processes are affected by the error-generating process to different degrees.

7 Results

Having established Model 2 as an adequate model of RC attachment in the context of question-answering, we used its parameter estimates to understand the pattern of responses in the experimental data: Figure 6 shows the Model 2 population parameter estimates for both experiments as well as 95% credible intervals for all four parameters. In addition to the difference in the guessing bias g between experiments, it also shows a lot of uncer-
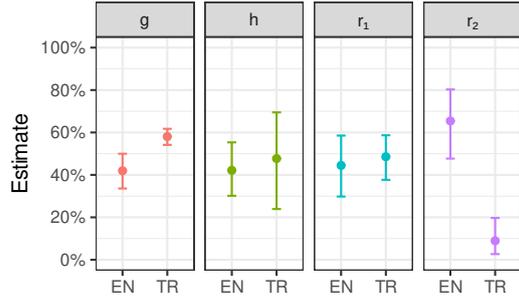
tainty in the estimates of the attachment parameter h, which represents the probability with which the parser adopts an N1 attachment interpretation over an N2 attachment structure in ambiguous attachment conditions. Both estimates of the N1 attachment probability have rather wide credible intervals, with 42% ($CrI = [30; 55]$) for the English experiment and 48% ($CrI = [24; 70]$) for the Turkish experiment. While the estimate for English is consistent with weak evidence for an N2 attachment preference, the estimate for Turkish indicates no clear preference.

The explanation for the surprising absence of evidence for an N1 attachment preference in the parameter h in Turkish lies in the the substantial difference between the successful recall probabilities r_1 (49%, $CrI = [38; 59]$) and r_2 (9%, $CrI = [3; 20]$), which indicate that N1 interpretations were successfully processed and recalled with a higher probability than their N2 counterparts. According to the assumptions of Model 2, this leads to a question response pattern which appears to suggest an N1 preference even when there isn't one ($h = 0.5$): When participants decide to adopt an N1 interpretation, their question responses indicate N1 attachment on most trials – sometimes due to successful recall of the N1 interpretation, and at other times as a result of guessing. When participants decide to adopt an N2 interpretation, however, they fail to recall the correct interpretation most of the time, and thus engage in guessing. Importantly, guesses result in N1 responses 50% of the time, since questions about N1 and N2 interpretations are balanced. As a result, a substantial difference between r_1 and r_2, such that $r_1 < r_2$ will lead to more N1 responses than N2 responses to questions about ambiguous sentences because

N1 interpretations are more successfully recalled, even if ambiguous sentences are assigned N1 interpretations only 50% of the time.

Whatever the source of higher error rates in the N2 attachment conditions in the Turkish experiment is, our MPT analysis suggests that what appears as a weak N1 attachment preference in our Turkish experiment is actually a consequence of a large number of guessing trials associated with N2 attachment. In sum, our analysis shows that (i) the N2 attachment preference in the English experiment appears to hold up even when guessing trials are taken into account, and (ii) that what appears to be an N1 attachment preference in Turkish is readily explained by the processing difficulty associated with processing and recalling N2 attachment structures in Turkish.

8 Summary

Based on the assumption that readers sometimes do not allocate the required amount of attention to the task they are performing, we have discussed a previously neglected source of bias and variability that may affect studies of attachment preferences and of interpretation preferences more generally. We attempted to account for the role of guessing as a strategy used in answering comprehension questions when the answer is not known. We argue that understanding the role of guessing in discrete choice tasks is crucial because data consisting of responses to comprehension questions where participants sometimes fail to arrive at a full interpretation of the structure may be confounded. To this end, we proposed an MPT-based analysis method that allows to de-confound parameters of theoretical importance from nuisance parameters such as the guessing rate. We tested two variants of the MPT-based model on experimental data from English and Turkish, and demonstrated that this method can provide further insight into the processes underlying participants' answering behavior as well as their attachment preferences.

References

William H. Batchelder and David M. Riefer. 1999. Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6(1):57–86.

Paul-Christian Bürkner. 2018. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411.

Fernando Cuetos and Don C. Mitchell. 1988. Cross-linguistic differences in parsing. *Cognition*, 30(1):73–105.

Fernando Cuetos, Don C. Mitchell, and Martin Corley. 1996. Parsing in different languages. In *Language Processing in Spanish*. Erlbaum, Hillsdale, NJ.

Brian Dillon and Matthew Wagers. 2019. Approaching gradience in acceptability with the tools of signal detection theory. Preprint, Open Science Framework.

Nazik Dinçtopal-Deniz. 2010. Relative clause attachment preferences of Turkish L2 speakers of English. In Bill VanPatten and Jill Jegerski, editors, *Language Acquisition and Language Disorders*. John Benjamins Publishing Company, Amsterdam.

A. Drummond. 2013. Ibex farm.

E. Erdfelder, T. Auer, B. Hilbig, A. Aßfalg, M. Moshagen, and L. Nadarevic. 2009. Multinomial Processing Tree Models: A Review of the Literature. *Journal of Psychology*, 217(3):108–124.

Lyn Frazier. 1987. Sentence Processing: A Tutorial Review. In Max Coltheart, editor, *Attention and Performance XII*. Erlbaum, Hillsdale, NJ.

E. Gilboy, J.-M. Sopena, C. Clifton, and L. Frazier. 1995. Argument structure and association preferences in Spanish and English complex NPs. *Cognition*, 54(2):131–167.

Nino Grillo, João Costa, Bruno Fernandes, and Andrea Santi. 2015. Highs and Lows in English Attachment. *Cognition*, 144:116–122.

Bilal Kırkıcı. 2004. The processing of relative clause attachment ambiguities in Turkish. *Turkic Languages*, 8:111–121.

K. McRae, M.J. Spivey-Knowlton, and M.K. Tanenhaus. 1998. Modeling the Influence of Thematic Fit in On-line Sentence Comprehension. *Journal of Memory and Language*, 38(3).

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

J. Smallwood. 2011. Mind-wanderingwhile reading. *Language and Linguistics Compass*, 5(2):63–77.

Stan Development Team. 2020. RStan: the R interface to Stan. R package version 2.19.3.

A. Staub. 2007. The return of the repressed. *Journal of Memory and Language*, 57(2):299–323.

B. Swets, T. Desmet, C. Clifton, and F. Ferreira. 2008. Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, 36(1):201–216.

A. Vehtari, Gelman A., and Gabry J. 2017. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27.