Prominence-sensitive pronoun resolution:

New evidence from the Speed-Accuracy Tradeoff procedure

Dave Kush
Norwegian University of Science and Technology
Haskins Laboratories

Clinton L. Johns
Haskins Laboratories

Julie A. Van Dyke
Haskins Laboratories

ADDRESS CORRESPONDENCE TO:

Dave Kush
Department of Language and Literature
Norwegian University of Science and Technology
NO-7491 Trondheim
Norway
dave.kush@ntnu.no

Abstract

Past studies have shown that antecedent prominence affects the processing of a pronoun, but these studies have used experimental methodologies that do not make it possible to determine at what stage(s) of pronominal resolution these effects occur. We used the Speed-Accuracy Tradeoff procedure to investigate whether antecedent prominence affects the accuracy of antecedent retrieval, the speed of resolution, or both. Consistent with previous results, we find that accuracy is higher when antecedents are prominent than when they are not (cf. Foraker & McElree, 2007). However, in contrast to previous results, we also find that prominence impacts the speed with which the pronominal dependency is resolved. We consider the implications of our findings for various models of pronoun resolution and offer suggestions for how to implement prominence-sensitive speed differences within a cue-based retrieval architecture.

Keywords: anaphora; prominence; speed-accuracy tradeoff; memory retrieval; sentence processing

Pronouns are anaphoric elements that often refer to entities introduced in some previous context. For example, in order to understand the short passages below, comprehenders are likely to interpret the pronoun *it* as coreferent with the antecedent noun phrase (NP) *the bouquet*.

(1) The bouquet was surprisingly fragrant. It smelled most of peonies.

(2) The florist delivered the bouquet. The widow found it on the doorstep.

Because of the distance between the anaphor and its referent in these examples, it is commonly assumed that interpreting *it* first requires retrieval of its antecedent from memory (e.g., McKoon & Ratcliff, 1980; Sanford & Garrod, 1998, 2005), after which processes related to resolution and semantic integration occur. There is ample evidence that an antecedent NP's agreement features (e.g., number, gender) are used to retrieve it from memory (Arnold, 2000; Foraker & McElree, 2007; Garrod & Terras, 2000; Sanford, Garrod, Lucas & Henderson, 1983; Sanford & Garrod, 2005; Stewart, Pickering & Sanford, 2000). However, it is less clear how other properties of the antecedent influence resolution. For example, in a well-structured, coherent discourse, pronouns are pragmatically constrained to refer back to entities that the local discourse segment is perceived to be 'about' (Ariel, 1990; Grosz, Joshi, & Weinstein, 1995; Gordon & Hendrick, 1997, 1998; Gundel, 1999). Although the terminology varies, such NPs are consistently characterized as *prominent* in the sentence or discourse representation (Gerrig & McKoon, 1998; Gordon, Grosz & Gillom, 1993; Gordon & Hendrick, 1997, 1998; Greene, McKoon & Ratcliff, 1992; Grosz et al. 1995; Gundel, 1999; McKoon, Gerrig, & Greene, 1996). Our goal is to examine how such prominence influences the resolution of pronominal anaphors.

NPs that occupy salient positions in their local syntactic environments, such as the matrix subject position (as *the bouquet* does in 1) are typically prominent (see, e.g., Engdahl &

Vallduvi, 1996; Gordon & Hendrick, 1998; Grosz et al., 1995). Many researchers have exploited this correlation between structural and discourse prominence to investigate the effects of prominence on pronoun resolution (for reviews see, Garnham, 2001; Garrod & Sanford, 1994). This work has shown that an antecedent's syntactic position influences the ease with which a pronoun is processed during reading: pronouns with prominent antecedents are read more quickly than pronouns with antecedents that are less prominent (e.g., Hudson, Tanenhaus & Dell, 1986). However, although prominent NPs enjoy an overall processing advantage as antecedents for pronouns, the source of this advantage – and its relation to retrieval – remains unclear. There are two aspects of antecedent retrieval that could be affected by cognitive prominence (Foraker & McElree, 2007). First, an NP's prominence could affect its *availability*, defined as the probability that the antecedent representation is accurately retrieved from memory. That is, prominent NPs might be successfully retrieved more often than less prominent NPs. Second, prominence could affect an NP's *accessibility*, which relates to the speed with which the target antecedent is identified and resolved. However, most previous research does not clearly distinguish between these two factors. One primary contribution of the current work is to employ a method – the Speed Accuracy Tradeoff method – that enables potential effects of prominence on antecedent availability and accessibility to be examined separately.

Most models of pronoun resolution are consistent with the idea that prominence affects antecedent availability. However, there is less agreement about the effect prominence may exert on accessibility. *Prominence-insensitive* models assert that NPs that agree with a pronoun in features like number and gender should be retrieved and interpreted equally quickly, irrespective of their cognitive prominence (Foraker & McElree, 2007; Sanford & Garrod, 2005). *Prominence-sensitive* models, on the other hand, propose that prominent NPs should be

considered more rapidly than their non-prominent counterparts. Prominence-sensitive models differ in the mechanisms by which the speed effect is achieved. Under some models, prominent NPs occupy a distinct memory store (*attentional focus*) that is consulted first during pronoun resolution (Rigalleau, Caplan & Baudiffier, 2004; Greene, McKoon & Ratcliff, 1992; Gundel, 1999). Others propose that previously-seen NPs are stored in a list, with potential antecedents ranked according to prominence, and retrieved in the order of their ranking (e.g. Gordon & Hendrick, 1997; Grosz et al. 1995). Finally, there are also models that admit the possibility that prominence information is used as a (highly-weighted) retrieval cue alongside lexical features like number and gender (Cunnings, Patterson & Felser, 2013). Most importantly for the current study, all of the prominence-sensitive implementations make the same prediction regarding accessibility: prominent antecedents will be processed more quickly than less prominent antecedents. This contrasts sharply with prominence-insensitive models, which predict that discourse factors like prominence should not impact processing speed.

Most of the experimental paradigms used to investigate anaphor resolution (e.g., probe verification, speeded grammaticality judgments, self-paced reading and eye tracking) cannot adjudicate between prominence-insensitive and prominence-sensitive models because they are unable to distinguish between effects of availability and accessibility. This is because dependent measures in such paradigms are susceptible to participant-specific response thresholds, such that any observed effects – faster reading times, for example – may be attributed to changes in the speed of processing (accessibility), the accuracy of retrieval (availability), or both (for extended consideration of this issue, see McElree & Dosher, 1993). A paradigm that avoids this limitation is the *Speed-Accuracy Tradeoff* (SAT) procedure, which provides orthogonal indices of processing speed and response accuracy within a single experimental task. It accomplishes this

by modeling the full time course of processing – for example, from prior to the formation of a syntactic dependency, through a period of retrieval, and continuing until well after dependency integration has occurred (e.g., Dosher, 1976; Reed, 1973; Wickelgren, 1977). This differs from traditional techniques that collect a single response (e.g., reading/reaction time) per trial per participant, in that the goal is to derive robust models of individual performance based on a high number of observations per participant. In the particular variant used in our study, referred to as multiple-response SAT (MR-SAT), 17 responses (each an acceptability judgment) are obtained from each participant on each trial, providing fine-grained empirical support for modeling the shape of the entire response curve, rather than a single response along an otherwise unspecified processing continuum. Critically, materials in SAT studies must be designed so that all responses occur at the end of the sentence, ensuring that participant judgments only reflect processes linked to the critical retrieval. In this way, the retrieval event of interest is isolated, and any dual processing that may arise due to comprehending language at the same time as making a response is avoided. Moreover, unlike standard dual-task paradigms (e.g. complex span tasks), there is no task-switching involved in order to make a judgment (apart from pressing a button while reading). Instead, the secondary task of making a judgment is dependent on the reading task, rather than standing in competition with it.

SAT modeling has been used to investigate a diverse array of linguistic phenomena (for reviews, see Foraker & McElree, 2011; McElree, 2015; McElree & Dyer, 2013), such as resolution of noun-verb dependencies (Johns, Matsuki, & Van Dyke, 2015; McElree, Foraker & Dyer, 2003; Van Dyke & McElree, 2011), comprehension of verb phrase ellipsis (Martin & McElree, 2008, 2009) and clausal ellipsis (Martin & McElree, 2011), figurative and coerced

expressions (McElree & Nordlie, 1999; McElree, Pylkkänen, Pickering, & Traxler, 2006), scalar implicature (Bott, Bailey & Grodner, 2012) and syntactic reanalysis (Martin & McElree, 2018).

Most relevant for the current project is a previous SAT investigation of pronoun resolution, the results of which were taken to support models of pronoun resolution in which accessibility is *not* influenced by antecedent prominence (Foraker & McElree, 2007). This study capitalized on research showing that clefting increases the syntactic prominence of an NP, ostensibly placing such entities in discourse focus (Gundel, 1999; Gundel, Hedberg & Zacharski, 1993). To test clefting's effects on retrieval accuracy and processing speed, Foraker and McElree (Experiment 1) had participants judge the acceptability of short passages like (3), in which the second sentence contained a pronoun (*It*) that had its antecedent (*the lead paint*) in a preceding sentence. The antecedent was either clefted and therefore prominent (3a), or embedded within a pseudo-cleft (3b), where it was assumed to be non-prominent.

(3a) It was *the lead paint* that annoyed the safety inspector. <u>It flaked/*grimaced</u>

(3b) The one whom *the lead paint* annoyed was the safety inspector. <u>It flaked/*grimaced</u>.

Foraker and McElree observed significantly higher asymptotic accuracy when the antecedent of the pronoun was clefted, but there was no evidence that the clefted NP was processed (i.e., identified and integrated) more quickly than its non-clefted counterpart. They concluded that prominence increases the strength (or activation) of an antecedent representation in memory, which in turn boosts the probability that the antecedent will be successfully retrieved. This conclusion was based on memory research demonstrating that increased asymptotic accuracy is associated with the strength of a memory representation (e.g., Dosher, 1979; Wickelgren, Corbett, & Dosher, 1980). Thus, in this study, prominence affected the *availability* of NP

antecedents for retrieval, but no evidence for effects on the *accessibility* of NP antecedents was found.

The conclusions in Foraker and McElree (2007) rely on the critical assumption that the cognitive prominence of the antecedent NPs differed across conditions. However, we suggest that this may not have been the case. There is considerable evidence that clefted NPs are cognitively prominent due to *contrastive* focus (Kiss, 1998; see also Chafe, 1976). Thus, the clefted NPs – *lead paint* in (3a) and *safety inspector* in (3b) – should be prominent due to focus. The *lead paint* was assumed to be non-prominent in (3b) because it was not clefted and was therefore not focused. However, there is reason to believe that even though *the lead paint* was not focused in (3b), it may still have been prominent because it was interpreted as the *topic* of the sentence. There is evidence suggesting that topical information is afforded prominence during coreferential processing (Anderson, Garrod, & Sanford, 1983; Clifton & Ferreira, 1987; Cowles, Walenski & Kluender, 2007; see also Grosz et al., 1995; Gundel, 1999). Crucially, topic and discourse focus are in complementary distribution in cleft sentences: that is, an NP that is not clefted, such as *lead paint* in (3b), must be the topic (or part of the topic; Cowles et al., 2007; Krifka, 2008). Thus, in Foraker and McElree's SAT study, it is possible that both critical NPs enjoyed some degree of cognitive prominence, one due to clefting and one due to topichood. This is also consistent with proposals that acknowledge multiple mechanisms for conferring prominence within a discourse, with some gaining prominence because of information structure, and others due to structural or syntactic methods of focusing information (Greene et al., 1992; Rigalleau, Caplan & Baudiffier, 2004; Grosz et al., 1995). Consequently, although Foraker and McElree clearly demonstrate that a contrastively focused NP is processed at the same speed as a topic NP, we believe they did not clearly assess the difference between prominent and non-

prominent NPs.

Evaluating potential effects of prominence on speed during pronoun resolution requires a manipulation that unambiguously places only a single NP in a cognitively prominent position. To achieve this, we manipulated whether an antecedent NP (*the bouquet*) was the topic of a context sentence as a way of determining its cognitive prominence. In order to make the antecedent NP a topic, we placed it in main subject position, as in (4a). To remove its topic status, we embedded the NP within a relative clause (RC), attached to the main subject, as in (4b). Importantly, unlike in Foraker and McElree (2007), topic and focus were not in a trading relation in our design.

(4a) *The bouquet* that the widow received that morning rested by the graves.

(4b) The widow that *the bouquet* was received by rested beside the graves.

In (4a), participants should analyze the antecedent NP *bouquet* as the topic, whereas *widow* should be the topic in (4b). These NPs appear in the main subject position, which is the default topic position in English (Givon, 1983; Grosz et al., 1995; Gundel et al. 1993, Lambrecht, 1994; Reinhart, 1982), and which is often used to connote topicality in experimental work (e.g., Fletcher, 1984; Lesgold, Roth, & Curtis, 1979). Previous studies have also demonstrated that the main subject position confers benefits to an antecedent during anaphor resolution (e.g., Gernsbacher & Hargreaves, 1988; Gordon et al., 1993; Stevenson, Crawley, & Kleinman, 1994). In contrast, the antecedent NP *bouquet* should not be highly prominent in (4b), as it is not the main subject, it occupies an RC-internal subject position typically reserved for backgrounded information (Lambrecht, 1994; Grosz et al. 1995), and there is no other reason to suppose that it bears focus. Our goal was to use the design in (4) to reassess whether prominence impacts the availability and accessibility of an antecedent NP during pronoun resolution.

Method

Participants. Twenty-four participants were recruited from the greater New Haven community. All participants were native English speakers that were enrolled in, or recently graduated from, a four-year college or university. Data from four participants were excluded from further analysis because these participants failed to reach a minimum standard for accuracy in at least one experimental condition (see below for details). Data from one other participant were removed due to non-monotonic response profiles. Analyses were conducted on data from the remaining eighteen participants (mean age 21.8; 13 female). The study was approved by the Yale University Human Investigation Committee.

Materials. We created thirty-two sets of experimental items following the design illustrated in Table 1. The full list of items is available in the Appendix. Experimental items were two-sentence passages, which crossed the factors PROMINENCE and SENTENCETYPE. The first sentence of the two sentence passages always contained a critical inanimate NP (e.g., *the bouquet* in Table 1). PROMINENCE determined the position of the antecedent NP. In *Prominent* conditions the antecedent NP was the main subject of the first sentence. In *Non-prominent* conditions the antecedent NP was the subject of a relative clause (RC) that was attached to the main subject, making it unlikely that it would be treated as a topic.

((TABLE 1 ABOUT HERE))

SENTENCETYPE had two levels: *Pronoun* condition and *Control* condition. In *Pronoun* conditions the main verb in the second sentence took the pronoun *it* as its object. In these conditions participants needed to establish coreference between the pronoun and the critical NP, which was its antecedent. *Control* conditions were identical to the Pronoun conditions except

that the critical object pronoun was replaced by a full, non-coreferential NP. Given the absence of an anaphor, there was no opportunity to establish coreference in the Control conditions.

The SAT analyses (described further below) are based on *d'* calculations, which control for response bias. Hence, we created both an acceptable and an unacceptable version of each experimental item. Acceptability was determined on the basis of the semantic fit between the main verb and its object in the second sentence. In Control conditions, this entailed evaluating the semantic fit between the verb and the full NP in the second sentence. Determining an item's acceptability in Pronoun conditions required evaluating the fit between the verb and the referent of the critical pronoun: the antecedent NP. In acceptable pronoun sentences the verb could take the antecedent NP as an object (e.g. *watered the bouquet*), but the verb could not take the antecedent NP as an object in *Unacceptable* sentences (e.g., #*comforted the bouquet*). The acceptability of a test sentence could not be determined on the basis of the collocation of the verb and the pronoun: all verbs could, in principle, take objects that could be referred to using the pronoun *it*. Thus, participants' ability to accept or reject test sentences was crucially predicated on their establishing coreference between the antecedent NP and the pronoun. Control conditions were designed so that the acceptability of verb-object pairings in the second sentence was counterbalanced within item sets. Verbs that led to unacceptability when paired with a pronoun (*comforted it*) were paired with an appropriate object in control conditions (*comforted the mourners*) and vice versa. This counterbalancing eliminated the possibility that participants could reliably anticipate the acceptability of an item in advance of the post-verbal region.

In all conditions (including foils and fillers described below), the critical response period was the final phrase (underlined in Table 1). In test sentences, the final phrase consisted of the pronoun and an adverbial modifier. Thus participants were forced to make a judgment about the

sentence at the point where the final verb was integrated with the coreferential pronoun. As the referent of the pronoun determined whether the sentence was acceptable or not, participants' responses coincided with the point of anomaly detection. We chose not to include the verb in the critical region so that the critical region was identical across all Pronoun conditions within an item. In Control conditions the final phrase consisted solely of the adverbial phrase without the non-coreferential object NP. As we discuss below, this entails that judgments in the Control condition were slightly delayed relative to the point in the sentence where the acceptable/unacceptable decision could be made.

We constructed four test lists. Each test list contained two conditions from every item set for a total of 64 test trials per list. Care was taken in constructing the lists so that participants could not predict the acceptability of one condition in the list based on the acceptability of the other. Test lists also contained 102 additional filler sentences of varying length and complexity.

Procedure. We used the multiple-response variant (MR-SAT) of the SAT procedure (McElree, 1993; Wickelgren, Corbett & Dosher, 1980; for psycholinguistic applications, see Foraker & McElree, 2007; Johns, Matsuki, & Van Dyke, 2015; Van Dyke & McElree, 2011). As in previous SAT studies, participants read sentences that were presented phrase-by-phrase in RSVP format, at a rate of 350ms/phrase. A series of 17 response tones (100 ms in duration, 1000 Hz) began 300 ms prior to the onset of the final critical phrase and continued every 350ms over a 5950ms response interval throughout which the critical phrase remained on the screen. This ensured responses throughout the entire time course of processing, from before antecedent retrieval until well after its integration with its verb. Test materials were randomized within a session and presented on a personal computer running E-prime (Schneider, Eshman, &

Zuccolotto, 2002), which recorded button presses and response latencies. Before each trial, participants saw a screen that reminded them which keys corresponded to an acceptable and an unacceptable judgment.

Participants pressed either of the response keys to begin presentation of the next item. When the tones began just prior to the critical region, they were instructed to press both response keys simultaneously (indicating uncertainty about the sentences' acceptability prior to reading the critical phrase), and then to indicate their acceptability judgment by continuing to press one of the two response keys in synchrony with the remaining tones. Participants were encouraged to make their judgment as rapidly as possible and were told that they could switch their judgment over the course of the response interval if their decision about the item changed.

Participants completed a 30-minute practice session on their first day of participation. At the beginning of the training phase participants were familiarized with the acceptability judgments that they would be asked to make in the experiment. Participants first used paper and pencil to judge the acceptability of practice sentences, some of which had simple agreement errors or missing constituents, and some of which were similar in structure to our test items and fillers. Participants received verbal feedback on whether their answers were correct. Immediately after this, they received training in how to respond in the SAT paradigm. This session trained them to (i) respond initially with both keys followed by eliminating responses on the dispreferred key once their decision was made; (ii) respond in time to the 17 tones and (iii) change responses over the course of the trial if desired. After the training session, participants took a short break and then completed their first test session. Participants were given regular breaks over the course of this and subsequent experimental sessions; during some of these breaks participants completed unrelated behavioral tasks for another study.

Data Analysis. We computed average response accuracy at each response point using a standard d' measure ($d' = z$(hits) – z(false alarms)), where z(.) is the inverse normal function.[1] A *hit* was defined as an "acceptable" response to an acceptable item, while a *false alarm* was an "acceptable" response to an implausible item. Accuracy in each acceptable condition was scaled against its corresponding unacceptable condition to create a discriminative d' score (MacMillan & Creelman, 2004). Participants whose asymptotic accuracy indicated near-chance performance on at least one experimental condition (i.e., $d' < 1$) were excluded. We calculated lag-latency by adding the average response time at each response tone to the latency of that tone. We plotted discriminative d' accuracy as a function of lag-latency in each condition and fit this response curve to a 3-parameter (λ, β, δ) exponential approach to a limit using the equation in (5). The three parameters determine distinct properties of the response curve: λ determines the asymptote of the curve and provides an estimate of participants' maximum discrimination accuracy when given sufficient processing time; β describes the rate of rise in accuracy from the intercept δ, where accuracy initially deviates from zero (i.e., chance).

(5) $\quad d' = \lambda\,(1 - e^{-(\beta(t-\delta))})$ for $t > \delta$, otherwise 0

Together, the rate and intercept parameter determine the overall temporal dynamics of processing. We were primarily concerned with testing whether there were reliable differences in the speed dynamics and we had no specific hypotheses about whether differences would manifest in rate or intercept.

[1] Speed-accuracy tradeoff analyses are typically conducted on *d'* measures, to avoid possible distortions of the shape of the functions caused by response biases. To verify that our results were not particular to the *d'* transform, we performed analogous fits on the proportion correct and false alarm data. The results show the same effects found in the *d'* analyses.

Our analysis of the data employed a hierarchical model selection procedure commonly used in SAT studies, including almost all previous investigations of linguistic processes (e.g., Foraker & McElree, 2007; Johns, Matsuki & Van Dyke, 2015; Martin & McElree 2008, 2009; McElree, 2000; McElree, Foraker, & Dyer, 2003; Van Dyke & McElree, 2011). The procedure is anchored by an analysis of the asymptotes from the participants' behavioral responses using mixed linear models implemented by lme4 (Bates, Mächler, Bolker & Walker, 2015) in the R statistical computing environment (R Core Development Team, 2016). All models included by-subject random intercepts. This analysis provides an important empirical analog for the subsequent analyses of the estimated asymptote parameter(s), which determines the number of unique asymptotes to include in the SAT model. Thus, following the empirical analysis, we consider the parameter estimates for a series of models of participants' response functions, beginning with the asymptote parameter, and then proceeding to the dynamics parameters (rate, intercept, speed composite). First, a "null" model (1λ-1β-1δ), in which all conditions are assigned the same parameter values, is fit to both the individual participant data and the average group data. Parameters are added iteratively until additions no longer improve model fit. The best fitting model for both each participant and the averaged data is determined according to two criteria: goodness-of-fit, and consistency of model fit across participants. We employed two goodness-of-fit measures: (i) the adjusted R^2, which measures the proportion of variance accounted for by the model after adjusting for its number of free parameters (e.g., McElree 1993); and (ii) the Akaike Information Criterion (AIC; Akaike, 1974), a selection measure that takes into account both traditional goodness-of-fit and model complexity. For the adjusted R^2 statistic, higher values indicate that a model closely fits the data. The opposite is true for the AIC statistic, such that models with lower AIC values are preferred. Although adjusted R^2 is

frequently assessed in published SAT studies, the AIC is not. However, the AIC provides an additional metric for model selection, and if both adjusted R^2 and AIC concur, this constitutes converging evidence for a particular model (Liu & Smith, 2009).

Parameter estimates were obtained using the *mrsat* package (Matsuki et al., *in preparation*) implemented in R.[2]The *mrsat* package identified the best-fitting model for each *d'* series by executing a fitting function that applies four different optimization algorithms 10 times each, with starting parameter values chosen at random for each run. The four algorithms were: (i) an iterative hill-climbing algorithm (Reed, 1976) akin to STEPIT (Chandler, 1969), implemented using the *acp* function, (ii) a limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm with box constraints (Byrd et al. 1995), implemented by the *optim* function, (iii) a box-constrained optimization algorithm based on PORT routines developed at Bell Labs (Fox et al., 1978), implemented by *nlminb,* and (iv) an unconstrained optimization algorithm based on a Newton-type method implemented in the *nlm* function (Dennis & Schnabel, 1983; Schnabel et al., 1985). The package compares the resulting parameter sets from each of the 40 runs, and selects the parameter set that best fits the data.

Results

Empirical asymptote data. SAT analyses typically begin with assessing potential differences in asymptotic accuracy, based on the average of the last four behavioral responses per condition. By considering the stable portion of the response curve, this procedure provides an empirical, data-driven anchor for subsequent modeling of the dynamics of the entire response function. Thus, we averaged *d's* for each participant's last four behavioral responses per condition (see Table 2). Although all conditions are presented together in Table 2, we analyze Pronoun and Control

[2] Available at GitHub (https://github.com/matsukik). Contact matsukik@gmail.com for details.

conditions separately, as the decision processes in the two comparisons are different and the two sets of conditions are not expected to share variance.

((TABLE 2 ABOUT HERE))

We observed a significant effect of Prominence on accuracy in both Pronoun and Control conditions. Average accuracy was higher in Pronoun-Prominent condition than in the NonProminent-Pronoun conditions (t = 7.10, p < .001; average d' difference = 0.91, 95% CI: 0.63 – 1.120). Similarly, average accuracy in the Control-Prominent condition was higher than in its Non-Prominent counterpart, (t = 2.53, p = .021), though the numeric difference was smaller than between Pronoun conditions (average d' difference = 0.39, 95%, CI: 0.01 – 0.77). These results suggest that, in our subsequent modeling analysis, data for both types of sentence will be best fit by a model with two asymptote parameters (2λ). Thus, we next followed a stepwise hierarchical modeling procedure, beginning by determining the optimal number of asymptote parameters for the data.

((TABLE 3 ABOUT HERE))

Hierarchical asymptote (λ) modeling. We began by fitting a 1λ-1β-1δ model, reflecting the null hypothesis that there are no differences between conditions, to both the average and individual participant data for both Pronoun and Control conditions. We then assessed whether a 2λ-1β-1δ model better fit the data in each.

Hierarchical asymptote modeling confirmed the results of our analysis of the empirical d' values, showing that two-asymptote models were warranted for both Pronoun and Control conditions. In Pronoun conditions the two-asymptote model better fit the average data (adjusted R^2 = 0.6743 v. 0.9939, see Table 3 for more details). A paired t-test on the estimated asymptote values confirmed that λ was significantly higher in the Pronoun-Prominent condition than in the

Pronoun-NonProminent condition (t = 7.04, p < .001). A two-asymptote model was also a better fit for the Control conditions (adjusted R^2 = 0.9544 vs. 0.9983), and estimated λ was also significantly higher in the Control-Prominent condition than in the Control-NonProminent condition (t = 2.50, p = .022). This pattern of results corresponds exactly to the observed effects of Prominence in the initial analysis of the empirical *d'* responses.

Hierarchical dynamics modeling. The second step of the hierarchical model selection procedure determined the optimal number of unique rate (β) and intercept (δ) parameters. For both Pronoun and Control conditions we compared three models of increased complexity to the baseline 2λ-1β-1δ model: a model with an extra rate parameter (2λ-2β-1δ), a model with an additional intercept parameter (2λ-1β-2δ), and a model in which both rate and intercept varied independently (2λ-2β-2δ).[3] Table 3 provides a quantitative summary of the full procedure.

Pronoun sentences. Estimated rates from individual participant 2λ-2β-1δ model fits were significantly faster in the Prominent condition than in the NonProminent condition (B = -0.84 (.33)*;* t = 2.54; p = .021).[4] Similarly, estimated intercepts from individual 2λ-1β-2δ model fits were significantly earlier in the Prominent condition than in the NonProminent condition (B = 0.122 (.05)*;* t = 2.34; p = .031). Although the 2λ-2β-2δ model exhibited the largest improvement over the 2λ-1β-1δ model on adjusted R^2 and 13 on AIC, neither individual rates nor intercepts

[3] Models reported in this section had asymptotes that were determined by the *mrsat* fitting function. Analyses were also run with corresponding fixed-asymptote models – models in which the asymptote parameter values for each condition were constrained to the empirical *d'* values for each condition – in order to minimize the likelihood that the dynamics differences that we observe are due to a parameter trade-off. The pattern of results in this analysis was identical.

[4] Closer inspection of the estimated parameters in the 2λ-2β-1δ model revealed that the model for one participant (S15) had an extreme rate estimate in the Pronoun-Prominent condition. Because exclusion of this participant's data did not affect the pattern of results – estimated rates were still significantly higher in the Pronoun-Prominent condition than in the Pronoun-NonProminent condition (t = 2.36) – we retained this participant in our analysis in the interests of power.

were significantly different ($t < 1$ for both comparisons) in this model. This likely reflects the trade-off between rate and intercept parameters known to occur in saturated models (for discussion see, e.g., Liu & Smith, 2009).

We believe that the most conservative interpretation of these results favors the 2λ-2β-1δ model (detailed in Table 4, depicted in Figure 1A) as the best fitting model for Pronoun conditions. This model provides a better fit of the average data than the 2λ-1β-2δ model according to both adjusted R^2 and AIC. In addition, the 2λ-2β-1δ model provided a better fit for a greater number of individual participants than did the 2λ-1β-2δ model. Although the goodness-of-fit statistics for the average data slightly favor the 2λ-2β-2δ over the 2λ-2β-1δ model, our analysis offered no statistical support for the additional dynamics parameters in the more complex model. Therefore, the 2λ-2β-1δ model is the most parsimonious and statistically robust for these data. (Once again, we note that our conclusions do not ultimately depend on whether the second speed parameter manifests on either rate or intercept, but on the evidence that an extra speed parameter is warranted at all.)

((FIGURE 1 ABOUT HERE))

((TABLE 4 ABOUT HERE))

Control sentences. We fit the average data and individual participants' data to the three more complex models described above and compared them to the base 2λ-1β-1δ model. In contrast to the Pronoun conditions, none of the more complex models offered consistent improvement over the baseline 2λ-1β-1δ model for the average data, either on adjusted R^2 or AIC. Rather, the more complex models were either equivalent or slightly worse than the baseline model, suggesting that their additional parameters were not warranted. At the participant level, additional dynamics parameters improved fits for some individual models, but the direction of

speed effects was inconsistent across participants. Paired t-tests confirmed that neither estimated rates nor intercepts differed reliably across conditions for any model ($t < 1$ for all comparisons). Thus, because there is no support for including additional dynamics differences between Control conditions, the 2λ-1β-1δ model best fits these data (detailed in Table 5, depicted in Figure 1B).

Before moving on, we wish to acknowledge one final difference between the Pronoun and Control conditions: intercepts for the Control conditions were much earlier than those for Pronoun conditions, suggesting that participants were able to determine whether Control items were acceptable or not more quickly than they were able to make a judgment in the Pronoun conditions. This likely reflects, at least in part, the fact that participants made their well-formedness judgments at different points in Pronoun and Control conditions. In Pronoun conditions participants gave a response immediately upon seeing the object of the verb (the pronoun), so the plotted function in Figure 1A provides a window of processing from the earliest stages of anomaly detection. However, in Control conditions participants made their decisions one word after the object (the full NP), giving them an extra 300ms of processing time before having to make their decision. While this effectively adds a constant to the response time in both conditions, it does not impact the critical effect of prominence because participants could assess the plausibility of both Control conditions equally easily by simply evaluating the relation between adjacent constituents (*comforted/*watered the mourners*).

General Discussion

Our study was designed to assess whether an antecedent NP's cognitive prominence affects its availability and accessibility during pronoun resolution. The MR-SAT paradigm allows separate assessments of processing speed and probability of retrieval (i.e., accuracy). Our manipulation of prominence allowed us to test whether the retrieval of a target antecedent was

more accurate when it was prominent and whether processing of the coreferential dependency was faster when the antecedent was prominent. This contrast allowed us to test the opposing predictions of two different accounts of antecedent retrieval, *prominence-insensitive* models and *prominence-sensitive* models, both of which predict that prominence should affect availability, but diverge as to whether prominence should impact accessibility. Prominence-insensitive models predict that an antecedent that matches gender and number features with a pronoun should be accessed at the same speed, regardless of prominence, because prominence is not counted as a retrieval cue. In contrast, prominence-sensitive accounts suggest that more prominent NPs are processed more quickly than those that are not. Our results clearly support the latter view.

Regarding availability, we observed the expected effect of prominence, with retrieval of prominent NPs being more accurate on average than retrieval of non-Prominent NPs. This suggests that Prominent NPs have higher baseline activation in memory (Nairne, 1990), a conclusion that aligns with that of Foraker & McElree (2007). It should be noted that we did observe some accuracy differences that cannot be attributed to antecedent availability's effect on retrieval: participants were also more accurate in the Control-Prominent condition relative to its NonProminent counterpart. This difference cannot stem from pronoun resolution, since participants did not need to process a pronoun in Control conditions. We suggest that this may relate to the processing of the RC in the first sentence. In our materials, the RC attached to the main subject was a non-subject RC. In Prominent conditions, the head of the RC was inanimate, whereas it was animate in NonProminent conditions. Previous studies have shown that participants have more difficulty processing the former RC type than the latter (Traxler, Morris & Seely, 2002, 2005; Gennari & MacDonald, 2008). Thus, the reduced accuracy associated with

processing NonProminent sentences can be partially attributed to the added cost of processing the RC. However, this animacy effect does not fully explain the asymptotic differences between Pronoun conditions, as there is a substantial residual *d′* difference between pronoun conditions even after the difference between control conditions is subtracted out (0.91 - 0.40 = 0.51). This difference, we believe, reflects the effect on activation associated with NP prominence.

We also observed that antecedent prominence does appear to affect accessibility: in the Pronoun conditions, faster processing rates were observed with prominent antecedents than embedded, non-prominent antecedents. In contrast, no reliable speed differences emerged between structurally-matched Control conditions in which processing did not require a coreferential relation to be resolved. This result is inconsistent with the findings in Foraker & McElree (2007), who found no support for the idea that prominence conferred any advantage in processing speed. We propose that this disparity reflects the fact that the two studies differ in how prominence was achieved across their experimental conditions. As discussed above, Foraker & McElree's materials simultaneously invoked two different means of conferring prominence: syntactic position and information structure ("topichood"). Because of this, both their clefted and non-clefted NPs may have been prominent. Thus, Foraker & McElree's results show only that two types of prominence (e.g. "focus" and "topic") do not differentially affect processing speed – but do not speak to potential processing differences elicited by prominent and non-prominent antecedents.[5]

[5] It is also possible that the effects in Foraker & McElree (2007) do not offer a clear picture of the dynamics of antecedent access due to an additional confound. The acceptability of test sentences was determined by an animacy manipulation between the pronoun and the following verb (*He grimaced/It flaked* v. **It grimaced/*He flaked*). This manipulation makes it possible for participants to judge sentence acceptability without needing to establish pronoun-antecedent relations. Thus, acceptability could reflect collocational or bigram frequency of the pronoun and the verb, rather than a decision that required resolution of the pronoun.

The difference in rate could have arisen for a number of distinct reasons, each with distinct theoretical implications. On one interpretation, the difference in rate between the Pronoun-Prominent and the Pronoun-NonProminent condition could reflect differences in retrieval latencies: the process of retrieving a prominent antecedent from memory could simply take less time than retrieval of a non-prominent antecedent. Certain models of retrieval in sentence processing would allow retrieval latencies to vary as a function of cue-match between a pronoun and its antecedent (e.g., the ACT-R parser of Lewis & Vasishth, 2005). Thus, if prominence is used as a cue during retrieval, this could result in the rate difference we observed. We note that the conclusion that rate differences reflect differences in the speed of the retrieval process itself is, however, at odds with other parsing models that assume that a single, time-invariant retrieval mechanism subserves the construction of diverse linguistic dependencies in sentence comprehension. Such a model has been motivated largely by the absence of dynamics differences in SAT studies of filler-gap processing (e.g., Johns et al., 2015; McElree et al., 2003) and ellipsis resolution (Martin & McElree, 2008, 2009). However SAT studies from the recognition memory literature suggest that retrieval speed can be modulated when order information must be considered (e.g., McElree & Dosher, 1993). Given that pronominal dependencies are subject to a number of position-based constraints that do not apply to the other dependency types discussed above (Chomsky, 1982; Lee & Williams, 2008; Chow, Lewis & Phillips, 2014), we consider it an open question whether the same time-invariance is a characteristic of antecedent retrieval during anaphor resolution (see also Dillon et al. 2014).

Speed differences can also be modeled without positing differences in retrieval latency. Popular models of pronoun resolution offer two avenues for explanation. One interpretation holds that speed differences follow from an architecture in which retrieval is not necessary for

identifying a prominent antecedent, but is necessary to identify a non-prominent antecedent. The alternative interpretation is that speed differences reflect processes associated with *reanalysis*. We consider each of these explanations and comment on their plausibility against the backdrop of current conceptions of the memory architecture underlying incremental sentence processing.

Some prominence sensitive models assume that the parser *actively* maintains prominent antecedent NPs in a special memory store, alternatively called 'discourse focus' or 'attentional focus' (among other things), while non-prominent antecedents are *passively* represented outside attentional focus (Greene, McKoon & Ratcliff, 1992; Gundel, 1999; Myers & O'Brien, 1998; Rigalleau & Caplan, 2000; Rigalleau, Caplan & Baudiffier, 2004). Items in the focus of attention are assumed to be immediately accessible for processing with no need for retrieval, whereas accessing items outside of focus involves retrieval, which incurs a temporal cost. Thus, under this account, prominent antecedents are identified and processed more quickly than non-prominent antecedents because prominent antecedents need not be retrieved for processing to begin. Such models are popular among many experimental psycholinguists and formal linguists, but contemporary models of memory suggest that they may not be theoretically viable. Specifically, although readers can (and do) actively maintain information, there is considerable evidence that the capacity of focal attention is extremely limited, perhaps only to a single item (Cowan, 1995; Jonides et al., 2008; McElree, 1998, 2001, 2006; McElree & Dosher, 2001). In order for the main subject from the first sentence in our stimuli to occupy focal attention at the time when the pronoun was encountered in the second sentence, the reader would have had to maintain the subject NP across the entire first sentence and most of the second sentence. This is unlikely given that prior work also shows that the processing of a single clause-sized constituent is sufficient to displace the contents of focal attention (Johns et al., 2015; McElree et al., 2003).

As such, this kind of active maintenance would only be possible if the parser has distinct attentional foci for storing chunks corresponding distinct linguistic levels of representation (e.g. discourse and syntactic foci; for an implementation of this idea, see Lewis & Vasishth, 2005), for which we know of no independent evidence.[6]

Rate effects could also be driven by reanalysis. According to this possibility, in the Pronoun-NonProminent condition, initial antecedent retrieval would erroneously retrieve the first-mentioned subject NP instead of the non-prominent target antecedent. Resolution would subsequently fail, prompting an additional attempt to retrieve the correct antecedent. This would result in longer processing time before coreference was established. Explanations of this sort have been offered for rate differences in the processing of other dependencies that require reanalysis (Bornkessel, McElree, Schlesewsky & Frederici, 2006; Martin & McElree, 2018).

Some prominence sensitive models assume that reanalysis is required because initial antecedent retrieval is biased towards retrieving a prominent NP even when that NP does not match the morphological features of the pronoun. This bias can be achieved in different ways: Under well-known models such as *Discourse Prominence Theory* (Gordon & Hendrick, 1997) and *Centering Theory* (Grosz et al., 1995) previously-seen NPs are stored in a dynamically updated list, ranked in order of prominence. Retrieval then involves *sequential consideration* of candidate antecedents in order of prominence: readers first attempt to resolve a pronoun by linking it with the most prominently ranked NP in the list, irrespective of that NP's morphological features. Less prominent NPs are only considered after a more prominent NP is rejected as a suitable antecedent. Thus, slower rates would arise under these theories due to a

[6] There are task-specific circumstances under which such distant items may be maintained in focal attention: the task must specifically encourage participants' "chunking" the distant information with the current information that is under active processing (McElree, 1998), or to explicitly maintain the distant information (McElree, 2001, 2006). These findings come from memory paradigms, rather than from language processing, and as such, their connection to studies of linguistic operations has not been assessed.

reanalysis-triggered serial search procedure. In light of evidence for the limited capacity of active memory (see above), these models do not seem viable if they assume that the prominence-ranked list is actively maintained. However, we note that sequential consideration can, in principle, be implemented in a system that stores *all* prior NPs outside of the focus of attention, requiring the retrieval of both prominent and non-prominent NPs alike.[7]

Bias towards retrieving a mismatching prominent NP over a matching non-prominent NP is also possible under content-addressable models where prominence and agreement features are probabilistic cues that are differentially weighted and applied simultaneously during the retrieval of an antecedent. Cue-weighting schemes have been proposed to account for retrieval preferences in thematic processing (Van Dyke & McElree, 2011), agreement (e.g., Dillon et al. 2013) and, to a lesser extent, pronominal processing (Cunnings et al., 2014). A speed advantage for prominent NPs could be achieved if prominence cues were weighted more heavily than agreement cues. In our Pronoun-NonProminent condition, where retrieval must 'choose' between a prominent non-matching NP on the one hand, and a non-prominent matching NP on the other, differential cue weighting would mean that retrieval would *more often* opt for the prominent NP first. On such trials, re-retrieval would be triggered. However, the nature of probabilistic retrieval would ensure that the non-prominent NP would still be retrieved first on some proportion of trials. As a result, the reduced rate in our NonProminent condition would reflect an average over a (smaller) number of trials in which the matching non-prominent antecedent was successfully

[7] For example, within cue-based retrieval models (e.g., Lewis, Vasishth & Van Dyke, 2006), it is possible to envision an encoding system that recapitulated prominence rankings using a series of (dynamically-updated) 'prominence cues'. Within such a system, sequential consideration would proceed as follows: reading a pronoun would trigger a retrieval using a cue that diagnostically identified the most prominent NP in memory (e.g. [Prominence-rank:1]), ignoring gender or number features. If that retrieval failed to yield an appropriate antecedent, the parser could retrieve again, cuing for the next NP in the prominence hierarchy (e.g. [Prominence-rank:2]), and so on. Thus, successful retrieval time for a NP with prominence rank *n* would scale with the time it took to complete *n* retrievals.

retrieved initially and a larger number of trials where initial retrieval failed and additional retrievals were necessary.

It should be acknowledged that models that attempt to exploit 'discourse prominence' cues must provide an explanation of how such cues are encoded and updated during incremental processing. Cue-based systems typically employ inherently static features to encode objects in memory. Feature values are set at encoding time and are usually thought to remain fixed. Inherent features, such as lexical gender or number, are easily accommodated by such a system, because an NP's lexical features rarely change as parsing progresses. However, encoding notions of discourse prominence requires more work from such a system, because an NP's (or referent's) prominence can and often does change across the run of a discourse. In order to accurately track the dynamic nature of discourse structure, features that represent prominence rank or concepts like *topic* or *focus* would have to be updated on a dynamic basis (e.g., whenever a new NP was introduced, or the discourse signaled a shift in relative prominence). It is possible to implement such dynamic updating through sequences of retrieval and feature-overwriting. If NP_1 bore a [topic] feature, for example, introducing a new topic would necessitate the retrieval of NP_1 so that its topic feature could be over-written. Dynamic cue-updating of this sort has been proposed for encoding other variable features relevant to anaphor resolution and quantifier scope, such as c-command (for extended discussion see Kush, 2013; Kush, Lidz & Phillips, 2015). This kind of dynamic updating can be implemented in cue-based parsing systems, but predictions of such models have yet to be assessed.

Conclusion

In this paper we used the MR-SAT procedure to test how antecedent prominence affects retrieval accuracy and processing speed during pronoun interpretation. We found evidence that prominence affected both accuracy and speed. Previous research has suggested that prominence can affect retrieval accuracy, but our results are the first to show unambiguous effects on speed of processing. Competing implementations of prominence sensitivity exist, but we favor the probabilistic cue model, as it appears most consistent with known constraints related to focus of attention, and the prior evidence in favor of cue-based retrieval in content-addressable memory (e.g. McElree et al., 2003; Martin & McElree 2009, 2011, 2018; Van Dyke & McElree, 2011). Specifically, we favor an interpretation of our speed effects according to which longer average processing times for non-prominent antecedents reflect a mixture of trials in which participants fail to retrieve the correct antecedent first and must re-retrieve, and trials where they successfully retrieve the non-prominent antecedent on the first attempt. However, future investigation is required to both test this possibility and to clearly distinguish among possible models of prominence-sensitive pronoun resolution.

Acknowledgements

This work was funded by the following NIH grants to Haskins Laboratories: R01 HD-073288 (PI: Julie Van Dyke) and NRSA HD-080331 (PI: Dave Kush). We wish to thank the three referees – and Andrea Martin in particular – for helpful suggestions and feedback during the review process. We also thank Brian Dillon and Kazunaga Matsuki for discussion, as well as Morgan L. Bontrager and Anne Stutzman for their invaluable assistance in creating experimental stimuli and running participants, and P. R. Nelson for support and inspiration.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Anderson, A., Garrod, S. C., & Sanford, A. J. (1983). The accessibility of pronominal antecedents as a function of episode shifts in narrative text. *Quarterly Journal of Experimental Psychology*, *35*(3), 427-440.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London: Routledge.

Arnold, J. E., Eisenband, J. B., Brown-Schmidt, S. & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, *76*(1), B13-B26. doi: 10.1037/e501882009-378

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*(1), 123-142.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190-1208.

Chandler, J. P. (1969). STEPIT-Finds local minima of a smooth function of several parameters. *Behavioral Science*, *14*(1), 81-82.

Chen, Z., Jäger, L., & Vasishth, S. (2012). How structure-sensitive is the parser? Evidence from Mandarin Chinese. *Empirical Approaches to Linguistic Theory: Studies of Meaning and Structure, Studies in Generative Grammar*, 43-62.

Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding*. Cambridge, MA: MIT press.

Chow, W-Y, Lewis, S. & Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Frontiers in psychology*, 5, 630. doi: 10.3389/fpsyg.2014.00630

Clark, S. E. & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, *3*(1), 37-60. doi: 10.3758/bf03210740

Clifton, C., Jr., & Ferreira, F. (1987). Discourse structure and anaphora: Some experimental results. In M. Coltheart (Ed.), *Attention and performance 12: The psychology of reading* (pp. 635-654). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Cowan, N. (1995). *Attention and memory: An integrated framework.* Oxford Psychology Series (No. 26). New York, NY: Oxford Univrsity Press.

Cowles, H. W., Walenski, M., & Kluender, R. (2007). Linguistic and cognitive prominence in anaphor resolution: Topic, contrastive focus and pronouns. *Topoi, 26*(1), 3-18.

Cunnings, I., Patterson, C. & Felser, C. (2014). Variable binding and coreference in sentence comprehension: evidence from eye movements. *Journal of Memory and Language*, *71*(1), 39-56. doi: 10.1016/j.jml.2013.10.001

Dillon, B., Chow, W. Y., Wagers, M., Guo, T., Liu, F., & Phillips, C. (2014). The structure-sensitivity of memory access: evidence from Mandarin Chinese. *Frontiers in psychology*, *5*, 1025.

Dillon, B., Mishler, A., Slogett, S. & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language, 69*(2), 85-103. doi: 10.1016/j.jml.2013.04.003

Dosher, B. A. (1976). The retrieval of sentences from memory: A speed-accuracy study. *Cognitive Psychology*, *8*(3), 291-310.

Dosher, B. A. (1979). Empirical approaches to information processing: Speed-accuracy tradeoff functions or reaction time – A reply. *Acta Psychologica*, *43*(5), 347-359.

Vallduví, E., & Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics*, *34*(3), 459-520.

Fletcher, C. R. (1984). Markedness and topic continuity in discourse processing. *Journal of Verbal Learning and Verbal Behavior*, *23*(4), 487-493.

Foraker, S. & McElree, B. (2007). The role of prominence in pronoun resolution: Active versus passive representations. *Journal of Memory and Language*, *56*(3), 357-383. doi: 10.1016/j.jml.2006.07.004

Foraker, S., & McElree, B. (2011). Comprehension of Linguistic Dependencies: Speed-Accuracy Tradeoff Evidence for Direct-Access Retrieval From Memory. *Language and Linguistics Compass*, *5*(11), 764-783.

Gao, L., Liu, Z., & Huang, Y. (2005). Who is *ziji*? An experimental research on Binding Principle. *Linguistic Science, 2,* 39-50.

Garrod, S., & Terras, M. (2000). The contribution of lexical and situational knowledge to resolving discourse roles: Bonding and resolution. *Journal of Memory and Language*, *42*(4), 526-544.

Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, *58*(2), 161-187.

Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, *27*(6), 699-717.

Gerrig, R. J., & McKoon, G. (1998). The readiness is all: The functionality of memory-based text processing. *Discourse Processes*, *26*(2-3), 67-86.

Givón, T. (1983). *Topic continuity in discourse*. John Benjamins Publishing Company.

Gordon, P. C., Grosz, B. J., & Gillom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, *17*(3), 311-347.

Gordon, P. C., & Hendrick, R. (1997). Intuitive knowledge of linguistic co-reference. *Cognition*, *62*(3), 325-370.

Gordon, P. C., & Hendrick, R. (1998). The representation and processing of coreference in discourse. *Cognitive Science*, *22*(4), 389-424.

Greene, S. B., Gerrig, R. J., McKoon, G., & Ratcliff, R. (1994). Unheralded pronouns and management by common ground. *Journal of Memory and Language*, *33*(4), 511-526.

Greene, S. B., McKoon, G., & Ratcliff, R. (1992). Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 266-283.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203-226.

Gundel, J. K. (1999). On different kinds of focus. *Focus: Linguistic, cognitive, and computational perspectives*, 293-305.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274-307.

Hudson, S. B., Tanenhaus, M. K., & Dell, G. S. (1986). The effect of the discourse center on the local coherence of a discourse. In *Proceedings of the 8th Annual Meeting of the Cognitive Science Society*, 96-101.

Johns, C. L., Matsuki, K., & Van Dyke, J. A. (2015). Poor readers' retrieval mechanism: efficient access is not dependent on reading skill. *Frontiers in psychology*, *6*, 1552.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual review of psychology*, *59*, 193.

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3-4), 243-276.

Kush, D., & Phillips, C. (2014). Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in psychology*, *5*, 1252. doi: 10.3389/fpsyg.2014.01252

Kush, D., Lidz, J. & Phillips, C. (2015). Relation-sensitive retrieval: evidence from bound variable pronouns. *Journal of Memory and Language*, *82*, 18-40. doi: 10.1016/j.jml.2015.02.003

Lambrecht, K. (1994). Information structure and sentence form: A theory of topic, focus, and the mental representations of discourse referents. Cambridge, U.K.: Cambridge University Press.

Lee, M-W., and Williams, J. N. (2008). *The Role of Grammatical Constraints in Intra-Sentential Pronoun Resolution*. London: London Metropolitan University/Cambridge University manuscript.

Lesgold, A. M., Roth, S. F., & Curtis, M. E. (1979). Foregrounding effects in discourse comprehension. *Journal of Verbal Learning and Verbal Behavior*, *18*(3), 291-308.

Lewis, R. L. (2000). Specifying architectures for language processing: Process, control, and memory in parsing and interpretation. *Architectures and mechanisms for language processing*, 56-89.

Lewis, R. L., Vasishth, V. & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, *10*(10), 447-454. doi: 10.1016/j.tics.2006.08.007

Li, X & Zhou, X. (2010). Who is *ziji*? ERP responses to the Chinese reflexive pronouns during sentence comprehension. *Brain Research*, *1331*, 96-104.

Liu, Z. (2009). The cognitive process of Chinese reflexive processing. *Journal of Chinese Linguistics*, *37*, 1-27.

Li, C., & Smith, P. (2009). Comparing time-accuracy curves: beyond goodness-of-fit measures. *Psychonomic Bulletin & Review, 16*(1), 190-203.

Martin, A. E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, *58*(3), 879-906.

Martin, A. E., & McElree, B. (2009). Memory operations that support language comprehension: evidence from verb-phrase ellipsis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1231-1239.

Martin, A. E., & McElree, B. (2011). Direct-access retrieval during sentence comprehension: evidence from sluicing. *Journal of memory and language*, *64*(4), 327-343.

Martin, A.E., & McElree, B. (2018). Retrieval cues and syntactic ambiguity resolution: speed-accuracy tradeoff evidence. *Language, Cognition, and Neuroscience.* doi: 10.1080/23273798.2018.1427877.

McElree, B. (1998). Attended and non-attended states in working memory: Accessing categorized structures. *Journal of Memory and Language*, *38*(2), 225-252.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 817-835.

McElree, B. (2006). Accessing recent events. *Psychology of learning and motivation*, *46*, 155-200.

McElree, B. (2015). Memory processes underlying real-time language comprehension. In D. S. Lindsay, C. M. Kelley, A. P. Yonelinas, & H. L. Roedinger III (Eds.) *Remembering: Attributions, processes, and control in human memory. Essays in honor of Larry Jacoby* (pp. 133-152). New York, NY: Psychology Press.

McElree, B., & Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General*, *122*(3), 291-315.

McElree, B., & Dosher, B. A. (2001). The focus of attention across space and across time. *Behavioral and Brain Sciences*, *24*(1), 129-130.

McElree, B., & Dyer, L. (2013). Beyond capacity: the role of memory processes in building linguistic structure in real time. In M. Sans, I. Laka, & M. K. Tanenhaus (Eds.), *Language down the garden path: The biological and cognitive basis for linguistic structure* (pp. 229-240). Oxford, U. K.: Oxford University Press.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111-123.

McElree, B., Foraker, S. & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67-91.

McElree, B., & Nordlie, J. (1999). Literal and figurative interpretations are computed in equal time. *Psychonomic Bulletin & Review*, *6*(3), 486-494.

McElree, B., Pylkkänen, L., Pickering, M. J., & Traxler, M. J. (2006). A time course analysis of enriched composition. *Psychonomic Bulletin & Review*, *13*(1), 53-59.

McKoon, G., Gerrig, R. J., & Greene, S. B. (1996). Pronoun resolution without pronouns: some consequences of memory-based text processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 919-932.

McKoon, G., & Ratcliff, R. (1980). The comprehension processes and memory structures involved in anaphoric reference. *Journal of Verbal Learning and Verbal Behavior*, *19*(6), 668-682.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, 18(3), 251-269.

Reinhart, T. (1982). Pragmatics and linguistics: an analysis of sentence topics. Bloomington, IN: Indiana U. Linguistics Club.

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science*, *181*(4099), 574-576.

Reed, A. V. (1976). List length and the time course of recognition in immediate memory. *Memory & Cognition*, *4*(1), 16-30.

Rigalleau, F., & Caplan, D. (2000). Effects of gender marking in pronominal coindexation. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(1), 23-52.

Rigalleau, F., Caplan, D., & Baudiffier, V. (2004). New arguments in favour of an automatic gender pronominal process. *Quarterly Journal of Experimental Psychology Section A*, *57*(5), 893-933.

Sanford, A. J. & Garrod, S.C. (1998). The role of scenario mapping in text comprehension, *Discourse Processes, 26*(2-3), 159-190.

Sanford, A. J., & Garrod, S. C. (2005). Memory-based approaches and beyond. *Discourse Processes*, *39*(2-3), 205-224.

Sanford, A. J., Garrod, S., Lucas, A., & Henderson, R. (1983). Pronouns without explicit antecedents?. *Journal of Semantics*, *2*(3-4), 303-318.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime: User's guide*. Psychology Software Incorporated.

Stevenson, R. J., Crawley, R. A., & Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, *9*(4), 519-548.

Stewart, A. J., Pickering, M. J., & Sanford, A. J. (2000). The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language*, *42*(3), 423-443.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*(1), 69-90.

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

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3): 247-263. doi: 10.1016/j.jml.2011.05.002

Wagers, M. W. & McElree, B. (2011). Memory for linguistic features: Evidence from the dynamics of agreement. Unpublished manuscript.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67-85.

Wickelgren, W. A., Corbett, A. T., & Dosher, B. A. (1980). Priming and retrieval from short-term memory: A speed accuracy trade-off analysis. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 387-404.