Reading comprehension involves adding simple and composite discourse referents to a mental model

Suhail Matar (s.matar@bcbl.eu)
Paloma Morcillo Ortega (p.morcillo-ortega@uea.ec.uk)
Manuel Carreiras (m.carreiras@bcbl.eu)
The Basque Center on Cognition, Brain and Langauge (BCBL)

Abstract

During reading, the mind continuously builds and updates a discourse model—a cumulative mental world representing the gleaned information. A key operator in this process is establishing novel discourse referents—entities in the model that can be picked out. For instance, 'She bought *wool*, *sponges* and *steel*' establishes three simple referents (italicized). By contrast, 'She bought *sponges of steel wool*…' establishes two simple referents forming a composite, itself referenceable (e.g., '…and glued *them* together.'). Here, in an online reading study, we target the cognitive basis of establishing simple and composite referents in the discourse model. Participants (n=43) read 72 five-sentence English stories sentence by sentence. The fourth, critical sentence featured: (i) three simple referents ('*wool*, *sponges* and *steel*'; $simple_3$); (ii) two simple referents ('*steel wool* and *sponges*'; $simple_2$), or (iii) two simple referents forming a composite referent ('*sponges of steel wool*'; composite). Crucially, across conditions, critical sentences had identical lengths and lexical items, while remaining sentences were fully identical. A true/false comprehension task followed each story. We hypothesized that additional referents would increase reading times (RTs) beyond syntactic, semantic, and lexical factors. Multiple regression analyses showed significant effects of adding simple ($RT_{simple3}$>$RT_{simple2}$) and composite ($RT_{composite}$>$RT_{simple2}$) referents. The effects appeared only on critical (but not on subsequent) sentences, possibly reflecting the cognitive operators of establishing, rather than maintaining, novel referents. Our findings pave the way for future work investigating hypotheses of hierarchical structure-building in mental discourse models.

Keywords: reading comprehension; discourse model; discourse referent; discourse processing.

Introduction

Real-life reading comprehension often involves processing and tracking complex information. To achieve this, the reading mind builds a mental representation in which to keep track of key information across a text—meaning, across an unfolding discourse. The mind builds this representation, called a discourse model (also a situation model), dynamically and continuously, updating it when needed as it encounters new information in the text (Glenberg et al., 1987; Zwaan & Madden, 2004; Richter et al., 2018).

Consider, for instance, this story fragment: '*The cook was washing some pots. They were so greasy that she had to use a sponge of steel wool.*' The two sentences introduce several *discourse referents* to the story (e.g., *the cook*, *pots*)—elements in the discourse model that can be picked out and referred to (Johnson-Laird & Garnham, 1980). For example, in the second sentence '*they*' refers to the pots and '*she*' refers to the cook. Prior work has shown that establishing a new discourse referent ('*some pots*') and accessing an already established referent ('*they*') are two distinct operators (Murphy, 1984) with different neural correlates (Coopmans & Nieuwland, 2020; Nieuwland et al., 2019). However, the behavioral cost of adding novel referents remains unclear.

Moreover, human languages allow for more complex ways of referencing (Wittenberg et al., 2021). For instance, suppose the story above continued with the following sentence: '*It made her job much easier.*' The pronoun '*it*' here refers not to simple referents like '*sponge*', '*steel*', or '*wool*', but rather to '*a sponge of steel wool*'—i.e., a *composite* entity. Indeed, in a visual-world paradigm, when hearing '*Put the cup on the saucer, then put that on the table.*', participants were more likely to interpret the pronoun '*that*' as referring to the composite '*the cup on the saucer*' rather than the simple referent '*the cup*', and to move both objects (Brown-Schmidt et al., 2005).

Put together, the literature suggests that the brain tracks simple referents, and that humans can interpret linguistic elements (e.g., demonstratives or pronouns) as referring to composite entities. We also know that establishing novel simple referents and accessing already-established ones are distinct operators. However, (i) What is the cognitive or behavioral cost of establishing novel referents? (ii) Does the reading mind establish composite referents in the discourse model, beyond simple referents? (iii) If so, do adding composite referents and adding simple referents to the discourse model constitute distinct cognitive operators?

To answer these questions, we conducted an on-line behavioral study in English, where participants read short narratives. In an online sentence-by-sentence reading paradigm, participants read five-sentence narratives sentence by sentence as we measured their reading times (RTs). There were three different conditions. Across conditions, all the sentences in a stimuli set were identical, except for the fourth, critical sentence (see Table 1). The critical sentence ended either with 3 simple referents ($simple_3$: '*wool*, *sponges* and *steel*'), 2 simple referents ($simple_2$: '*steel wool* and *sponges*') or 2 simple referents that form a composite referent (composite: '*sponges of steel wool*'). See *Methods* for an explanation on why we consider 'steel wool' a simple discourse referent. Crucially, the critical sentences in a set had identical numbers of words and featured identical lexical items.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Table 1: Example stimulus set. Sentence 4 ending with a critical zone (bolded) that adds to the discourse model: three simple referents (simple$_3$), two simple referents (simple$_2$), or two simple referents and one composite referent (composite).

Position	Sentence		Condition
1	John visited the new abstract art exhibition yesterday afternoon.		
2	On display were many innovative and original art pieces.		
3	He saw a painting made of vivid colors and swirling shapes.		
4	He also saw a sculpture made from	**wool, sponges and steel.**	simple$_3$
		steel wool and sponges.	simple$_2$
		sponges of steel wool.	composite
5	It was a very popular exhibit with quality pieces.		

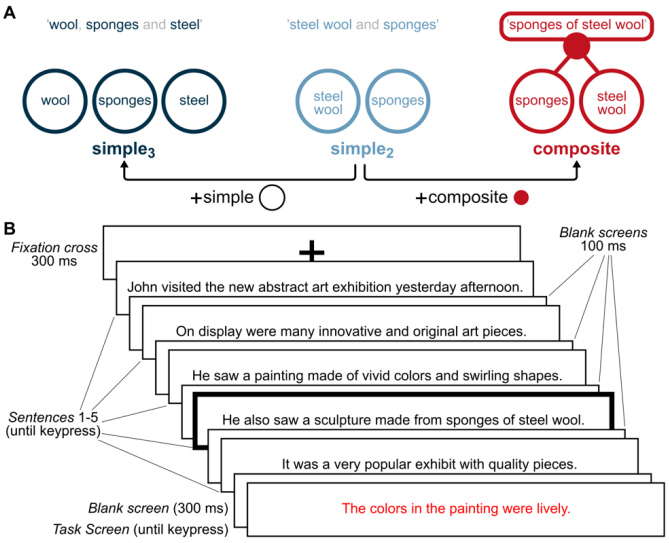

Figure 1: (a) experimental manipulation and (b) trial design.

Compared to simple$_2$, simple$_3$ has one additional novel simple referent (Fig. 1a). Therefore, we hypothesized that RTs would be longer in simple$_3$ compared to simple$_2$, reflecting the cost incurred by adding a simple referent—a +simple operator. Conversely, compared to simple$_2$, composite has one added, composite referent (Fig. 1a). Hence, if the mind indeed adds composite referents to the discourse model, we expected the cognitive operator +composite to cause longer RTs in composite versus simple$_2$. Finally, if introducing a simple referent and a composite referent are mentally distinct operators, this may lead to significantly different RTs between simple$_3$ and composite.

We test RTs on the critical sentence and the subsequent, final sentence. Prior research shows that discourse effects in reading are sometimes delayed (Duff & Altshuler, 2025; Frazier & Rayner, 1982; Rayner et al., 1992). We also hypothesized that effects on the critical sentence would reflect the cost of applying the cognitive operators +simple and/or +composite; conversely, we hypothesized that effects on the final sentence (identical across conditions) would reflect the cost of maintaining more elements in working memory (Schuh et al., 2016).

Importantly, evidence for the mind building composite referents in the discourse model would provide grounds for hypotheses in which situation models feature hierarchically organized discourse referents (see *Discussion*).

Methods

Participants We recruited 49 monolingual English-speaking participants via Prolific, who fulfilled the following pre-screening requirements: they spoke English as a mother tongue, were U.S. nationals, were located in the U.S., were between 18 and 35 years old, and had no language disorders nor literacy difficulties. Participants self-reported normal or corrected-to-normal vision.

The Ethics Committee of our institution approved the experiment, which we run online via Gorilla in accordance with the relevant regulations. Participants provided informed consent and received compensation. One participant failed to complete the experiment and was excluded, leaving 48 participants (24 female, 1 undisclosed; mean age=28.5, SD=4.1, range=20–35 years).

Materials. We created 72 sets of experimental stimuli across 3 conditions (see Table 1 for an example). Each set consisted of a 5-sentence story, which was identical across conditions, except for the fourth sentence (see *Introduction* and Fig. 1a). Within a set, all critical sentences had the same lengths and lexical items; the remaining sentences were fully identical.

To ensure the same lexical items across conditions in the critical sentence, we relied on two-word simple referents like '*steel wool*'. These two-word collocations appear in the simple$_2$ and composite conditions, while the simple$_3$ condition featured the same two words in an inverted order and separated by an intervening word ('wool, *sponges* and steel'), to encourage parsing the two words ('wool, […] and steel') as distinct, independent simple referents, and discourage their association ('*steel wool*'). We determined that referents like '*steel wool*' are discursively simple using a coordination diagnostic test with an intervening noun: we substitute the two-word referent into a test sentence that pries the two words apart, and assess ourselves whether the resulting sentence allows a reading that preserves the original concept. For example, the test sentence '*Hind uses steel and cotton wool to produce her art.*' cannot be taken to mean '*She uses steel wool* (and cotton wool).': separating the two

constituents breaks down what '*steel wool*' refers to. (By contrast, the test sentence can easily have the interpretation '*Hind uses steel, and she uses cotton wool*'). In comparison, '*Hind champions animal and human rights.*' does carry the interpretation '*She champions animal rights* (and human rights).', so we avoided phrases like 'animal rights' as two-word referents in our stimuli.

Prior work suggests discourse models are more reliably updated when the information is factual (e.g., '*Hind visited*') and free of hypotheticals or conditionals (e.g., '*Hind might visit*' or '*if Hind visits*'; Tulling et al., 2021). Thus, to ensure participants engage cognitive operators that update the discourse model, we composed stories using factual language, such as things that a character did or perceived.

Using a Latin square design, we created 3 lists of stimuli, each with 72 trials; each list featured only one condition per set. Each participant saw only one list, and only one condition per set (24 trials per condition).

Task After each story, participants read a task statement and judged its veracity based on the story (see Fig. 1b). Half the task statements were true; half were false. The veracity of the statement was based, with equal probability, on any of the five sentences. In the example in Table 1, based on sentence 3, the task item '*The colors in the painting were lively.*' is true. The correct answer was always independent of the experimental condition, so participants did not have to pay special attention to the critical zone.

Experimental procedure First, we asked participants to report average reading hours per day (range: 0–10, 0.5 hour increments); we used this as a nuisance regressor.

Then, a series of screens explained the study and the task, after which participants trained on 3 example stories. The training trials followed the same structure as our experimental trials. Following each trial we explained why the task statement was true or false.

The main experiment then started. We divided the 72 trials into 6 blocks, each with 12 experimental trials and two task-check trials (see below). Blocks and trial order within blocks were randomized per participant. After each block, a screen informed participants of their percent accuracy, and they could rest for a few moments before continuing.

Task-check trials were 1-sentence stories (e.g., '*Kat wanted to watch a new movie, so she went to the cinema.*') with the same type of task ('*Kat decided to go to a bookstore.*', which is false). They served to break the monotony of 5-sentence stories, forcing participants to pay attention from the get-go. We also used them to test whether experimental trials were cognitively taxing on the memory. The rationale was that, if experimental trials were considerably more demanding than check trials, then experimental trial accuracy would be significantly worse.

Trial structure is described in Fig. 1b. First, a fixation cross appeared for 300 ms, indicating trial beginning. Next, the first sentence appeared, and participants had to press the spacebar to proceed. A blank screen appeared for 100 ms after each sentence; this allowed participants to saccade to the left side of the screen in anticipation of the next sentence; it also avoided a brisk, unnatural visual transition between successive sentences. In the experimental, 5-sentence trials, the next four sentences followed suit: the sentence first, until key-press, then a 100 ms blank screen. After the last blank screen, the task item appeared until participants pressed answered. A final blank screen appeared for 300 ms.

Experimental stimuli appeared in black against a white background. Task statements appeared in red against a white background, to differentiate them from the trials themselves.

Syntactic metrics We added syntactic complexity variables to account for RT differences related to syntactic properties. This is important because our three conditions map onto different syntactic structures. But because syntactic structures varied freely across sets and sentences, we quantified structure complexity for all stimuli to use as nuisance regressors. We calculated 11 different measures of syntactic complexity. Four of those are unidimensional scores (see Agmon et al., 2024, for an overview): the Yngve score (Yngve, 1960), the Frazier score (Frazier, 1985), the Frazier-Roark score (Agmon et al., 2024), and the Mean Dependency Distance (Haitao Liu, 2008). In addition, we calculated seven scores (proposed as multi-dimensional vectors) derived from syntactic structures (Agmon et al., 2024): number of total clauses, number of relative clauses, average dependency distance of relative clauses, number of center-embedded structures, maximum depth of center-embedded structure, average number of words in a noun phrase, and the total number of adjectival and adverbial phrases. We applied a Principal Component Analysis (PCA) over the 11-dimensional vector describing syntactic complexity of every sentence, then extracted the top 3 components, which capture 86.1% of the variation in the original vectors. We used these as nuisance regressors.

GPT-2 surprisal To estimate sentence predictability, we calculated average surprisal values using GPT-2 (Radford et al., 2019), a transformer-based language model pretrained on a large corpus of English text. Each sentence, punctuation included, was tokenized and passed through GPT-2, and surprisal was computed as the negative $\log_2$ probability of each token given its preceding context. We averaged these values, yielding a mean surprisal score per sentence in bits, reflecting the model's estimate of processing difficulty based on lexical and syntactic predictability.

Semantic composition effort Beyond discourse-level differences, the critical zone in the critical sentence (see Table 1) also differs in terms of the semantic composition required. To quantify and account for these differences, we used sentence-transformer embeddings (Reimers & Gurevych, 2019) and computed the Euclidean distance between the critical zone phrase embeddings and the normed vector-sum of constituent noun embeddings. For example,

Table 2: Results of regressing z-transformed log(RTs) against two nested mixed-effects models. Bolded predictors are only included in the full model. Categorical predictors are followed by parentheses indicating the base level and the comparison level. *p*-values are calculated using Satterthwaite's formula for approximating degrees of freedom.

	Null Model			**Full model**		
Predictors	β	t-stat	*p*-value	β	t-stat	*p*-value
Intercept	-0.17	-0.88	0.388	-0.24	-1.25	0.219
Age	0.03	0.29	0.774	0.03	0.29	0.774
Sex (F vs M)	0.14	0.78	0.442	0.14	0.78	0.442
List (1 vs 2)	0.31	1.38	0.176	0.31	1.38	0.175
List (1 vs 3)	0.15	0.68	0.500	0.15	0.68	0.500
Hours reading	-0.07	-0.70	0.487	-0.07	-0.70	0.486
Sentence length	0.20	13.85	**<0.001**	0.20	13.73	**<0.001**
Prior sentence log(RT)	0.02	15.78	**<0.001**	0.02	15.89	**<0.001**
Syntax 1	0.005	0.37	0.710	0.004	0.28	0.781
Syntax 2	-0.03	-2.07	**0.038**	-0.02	-1.66	0.098
Syntax 3	-0.03	-2.25	**0.025**	-0.02	-1.98	**0.048**
Average GPT-2 surprisal	0.13	9.26	**<0.001**	0.12	9.12	**<0.001**
Composition effort	-0.002	-0.17	0.869	-0.02	-1.30	0.193
Position (4 vs 5)	-0.14	-6.31	**<0.001**	-0.02	-0.55	0.579
Position × Composition effort	0.06	3.28	**0.001**	0.10	4.58	**<0.001**
Condition (simple$_2$ vs composite)				0.10	3.05	**0.002**
Condition (simple$_2$ vs simple$_3$)				0.11	2.77	**0.006**
Position × Condition (simple$_2$ vs composite)				-0.15	-3.27	**0.001**
Position × Condition (simple$_2$ vs simple$_3$)				-0.20	-3.70	**<0.001**

for '*wool, sponges and steel*', we subtracted from the phrasal embedding (punctuation included), the vector-sum of the embeddings of '*wool*', '*sponges*', and '*steel*'. For '*sponges of steel wool*' and '*steel wool and sponges*', we repeated this process, but independently also subtracted from the phrasal embeddings the vector-sum of the embeddings of '*steel wool*' and '*sponges*'; we then picked the smallest of the two results, reflecting a "minimum effort" strategy. We hypothesized that greater distances between phrase and word embeddings reflected increased efforts to compose them semantically. We included composition effort as a predictor, with an interaction with factorial sentence position, to allow for differences between compositional processing during (sentence 4) or after (sentence 5) reading the critical zone.

Statistical analyses We performed analyses in RStudio using R (version 4.4.0) and the *lme4*, *lmerTest*, and *emmeans* packages. We removed 6 participants' data: one had incomplete data, and 5 achieved accuracies below 75% on the experimental trials and/or the task check trials (Fig. 2). This left us with data from 43 participants. We also removed datapoints belonging to one stimulus set (one trial per participant), due to a technical error. Finally, we removed incorrectly answered trials, as this could suggest inattention; as a result, we lost 11.86% of the data at this stage.

In all analyses, z-transformed log-transformed RTs were the dependent variable. In our main analysis, we built two nested linear mixed-effects models against which we regressed RTs. Fixed effects are specified in Table 2. The null model included these nuisance fixed effects: intercept, z-transformed variables (age, sentence length in words, daily reading hours, syntactic complexity, surprisal, composition effort, log(RT) of previous sentence) and categorical variables (sentence position, sex, experimental list), in addition to interaction between sentence position and composition effect. To this null model, we compared a full model which had the exact same regressors, with the addition of a fixed effect for our experimental manipulation (Condition): a 3-level categorical factor, dummy-coded with simple$_2$ as the reference level. We also added an interaction term between Condition and sentence position (2-level categorical: sentence 4 vs. 5), to test whether the effects appear only on the critical sentences. Both models included random intercepts per participant and per stimulus set, reflecting individual variations between participants and between reading demands of different sets. The difference between the full and the reduced model is thus one binary regressor reflecting the effect of adding a simple discourse referent (+simple: simple$_3$ vs. simple$_2$), another binary regressor reflecting the effect of adding a composite discourse referent (+composite: composite vs. simple$_2$), and two interaction terms between these binary regressors and a binary regressor of sentence position (sentence 4 vs. 5).

We regressed RTs against both models, then compared model fits using a likelihood ratio test. This establishes the improvement in fit that is uniquely attributable to our experimental manipulation. To understand the comparisons driving any effects, we also conducted pairwise comparisons between our three conditions per sentence position, correcting for multiple comparisons using the Tukey method.

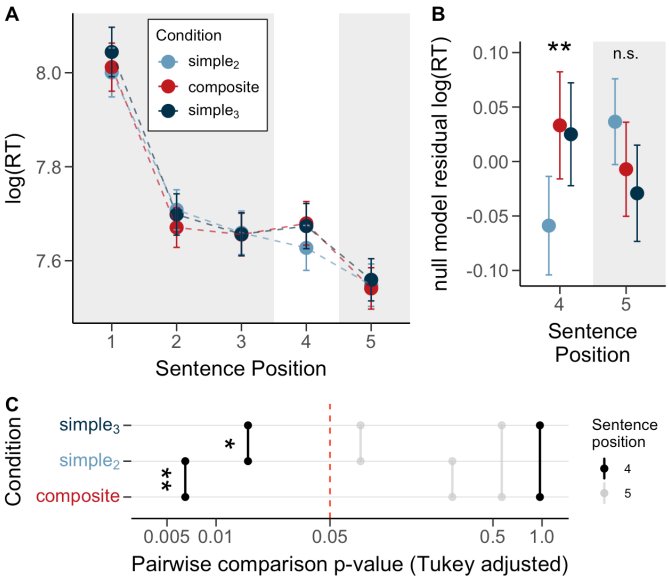

Figure 2: log-transformed RTs and referent effects. A) log(RTs) per sentence and condition, adjusted for cross-participant variation. Circles are adjusted mean RTs across participants; whiskers indicate 2.5th and 97.5th quantiles. White window highlights the critical sentence. B) residual log(RTs) of the null model in the critical (Sentence 4) and final (Sentence 5) sentences. Values are on z-transformed log-transformed scale. Circles indicate mean RTs across trials per sentence and condition. Asterisks indicate the effect of adding Condition to the null model. C) Tukey-adjusted p-values for pairwise comparisons between conditions in Sentence 4 (black) and 5 (gray). The end points of each bar indicate the conditions compared. The x-axis is a log scale.

Results

Accuracy Participants' mean accuracy on experimental trials was 86.4% (SD=7.68%); mean accuracy on task-check trials was 88.5% (SD=9.67%). A paired sample t-test between participants' accuracies on experimental and check trials revealed that we cannot reject the null hypothesis that they are drawn from the same distribution ($t(47)=1.657$, $p=0.104$). Based on accuracy results, we excluded data from 5 participants who scored below 75% on either trial type.

Reading time results We log-transformed RTs from 43 participants for statistical analysis. Fig. 2a shows adjusted grand average RTs per condition and sentence.

Next, we regressed the data against two nested models (see Table 2 and *Methods*). The null model represents the null hypothesis that our experimental Condition does not explain unique variance in reading times. The full model is identical to the null model, with the addition of the experimental *Condition*—as a main effect and as an interaction with sentence position. Results from both models show main effects of sentence length, syntactic complexity, average surprisal, and RT of prior sentence, in addition to interaction between composition effort and sentence position. The full model also showed significant interactions between *Condition* and sentence position, and a main effect of *Condition*.

A likelihood ratio test showed that adding *Condition* predictors significantly improves the full model's fit to the data (compared to the null model; $\chi^2(4)=17.36$, $p=0.0016$).

Next, we estimated the marginal means for each condition in each sentence position (4 or 5), and performed three pairwise comparisons separately per sentence position. The results, adjusted to account for multiple comparisons, appear in Fig. 2c. They show a significant effect of establishing a composite referent ($simple_2$ vs composite, $p=0.0065$) and of introducing an additional simple referent ($simple_2$ vs $simple_3$; $p=0.0157$) only on the critical sentence.

In RT terms, the estimated marginal means in the critical sentence are $\log(RT_{simple2})=7.70$, $\log(RT_{composite})=7.76$, and $\log(RT_{simple3})=7.76$. As all analyses were performed on log-transformed RTs, this translates to a factor of 1.062 for adding a composite or simple referent (equivalent to a 137 ms difference in raw RTs at the marginal means).

To further investigate the difference between adding composite and simple referents (composite vs $simple_3$), we conducted two one-sided tests (TOST) for statistical equivalence between the two conditions (Lakens, 2017). This allows us to test whether the difference between conditions is smaller than the smallest effect size of interest. Considering the online nature of the experiment and the potential lag differences between machines, we conservatively set the smallest effect size of interest to 50 ms. The results show that we cannot deduce that any pair of conditions are statistically equivalent. Thus, RTs for $simple_3$ and composite conditions are neither statistically equivalent nor different ($p=0.468$). In other words, it remains inconclusive whether building a composite referent and adding a simple referent are cognitively equivalent processes.

No effect of referent order A potential confound is referent order: Whether the two-word referents ('*steel wool*') appear first or second in composite and $simple_2$. There is some imbalance in our materials between the number of composite trials in which the two-word referents appear first or second (41 vs 30, respectively; i.e., 30 vs 41 in the $simple_2$ condition). To ensure that this does not affect the results, we conducted additional tests. First, we randomly sampled 60 sets from the total 71 sets: in 30 sets, the two-word referents appeared first in composite trials, and in the other 30 sets they appeared second, creating a fully counterbalanced subset. Next, using a likelihood ratio test to compare models as explained above, we independently tested the effects of adding either our experimental manipulation or the referent order (3-level factor: one-two, two-one, or $simple_3$) to the null model, extracting the p-value for each comparison. We repeated this process 1,000 times, randomly sampling the included sets each time, to ensure that the results were not due to the random sampling. This resulted in 1,000 *p*-values for every comparison. Referent order does not significantly explain RTs in the critical sentence in any of the 1,000 tests (all p-values > 0.05). Conversely, our experimental

manipulation explains significantly more RT variance when added to the null model in 99.8% of the tests.

Discussion

In this study, we set out to explore whether in reading comprehension the mind establishes composite referents in the discourse model. We also sought to quantify the behavioral cost associated with adding simple or composite referents to a discourse model. Finally, we asked whether adding extra simple (+simple) or composite (+composite) referents are cognitively distinct operators. To answer these questions, we used a simple, online sentence-by-sentence reading study, where the critical sentences varied in the number and type of novel discourse referents established.

Our main novel finding is that $RT_{composite} > RT_{simple2}$ (Table 2, Fig. 2)—i.e., that reading two simple referents (e.g., '*sponges*', '*steel wool*') that form a composite referent ('*sponges of steel wool*') incurs significantly longer RTs than reading the same two referents in coordination ('*steel wool and sponges*'). We also found that introducing one additional simple referent to the mental model incurs an increase in RTs ($RT_{simple3} > RT_{simple2}$; Table 2, Fig. 2). These effects appear even after controlling for other predictors, including syntactic, lexical, and semantic properties (Table 2). Controlling for referent order does not alter the results. Interestingly, the two effects appear only on the critical sentence, and not on the subsequent sentence (Fig. 2). These critical sentence effects likely reflect the demands of applying discourse-level mental operators (+composite or +simple) to update the discourse model, rather than the memory demands of maintaining the model over time (i.e., keeping the tally of objects in working memory). Thus, we interpret these effects as the costs of adding a composite or a simple referent, respectively, to the discourse model. Note that we do not imply here that there is no cost associated with maintaining the discourse model, only that the measured effects likely reflect the cost of updating the model. Participants were asked to pay attention and answer comprehension questions, not to explicitly encode the referents, so they may not have tried to actively maintain the objects in working memory.

Prior work found RT differences between establishing a new referent and accessing an already-established referent (Murphy, 1984) (e.g., a [new] truck versus the [previously talked-about] truck). Here, due to identical sentence lengths and lexical items across conditions, we are able to distill the effect of establishing a new referent on RTs (+simple). We note that, as in Murphy (1984), our participants read whole sentences, one by one. We concede that this results in a low-dimensional dependent variable. However, we hope to have shown that it is a simple, cheap and naturalistic metric, and that using it allows us to replicate and expand on work on discourse-level processing. This is especially true with the advent of large language models, which can help account for nuisance syntactic, semantic, and lexical factors.

To create our stimuli, we used two-word simple referents like '*steel wool*' (see *Methods*) that we identified using a coordination diagnostic test and our own intuition (e.g., '*Hind used steel and cotton wool.*' cannot mean she used steel wool and, independently, cotton wool; it can mean that she used cotton wool and steel, or perhaps a hybrid wool). One reviewer respectfully disagreed with our interpretation of the diagnostic test, but agreed with us that '*steel wool*' is a simple referent, as steel is an attribute of wool.

An important difference between $simple_3$ and the other two conditions is the presence of a comma in the critical zone (see Table 1), which can greatly influence reading times. However, both GPT-2 surprisal and the Composition Effort metrics were calculated using tokenized versions of our sentences, which include the comma tokens. This means that, these metrics account for the presence of commas or other punctuation in our stimuli, regardless of where they appear (say, which sentence position), just as they account for the presence of lexical tokens.

We believe that our findings open the door for a crucial shift in studying discourse-level processing during naturalistic language comprehension. Concretely, we show here that comprehension involves not only identifying simple discursive entities, but also stringing them together into composite referent units. This invites the hypothesis that building a mental discourse model consists of building a hierarchical referent structure. For example, in the sentence '*The artist bought a packet of sponges of steel wool.*', the composite referent '*a packet of sponges of steel wool*' is built of a simple referent '*packet*' and another, nested composite referent '*sponges of steel wool*'. Both composite units are referents because they can be picked out in the discourse model: in a continuation sentence like '*She opened it…*' the last pronoun refers to the '*packet of sponges of steel wool*', whereas in a continuation sentence like '*She used them…*' the last pronoun refers to the '*sponges of steel wool*'.

Conclusion

In this study, we set out to investigate whether there is a mental operator associated with building composite discourse referents from simpler units. To that end, we designed an online sentence-by-sentence reading study in English and found that reading two simple referents that build a composite, or reading three simple referents, incurs longer reading times than simply reading two referents, after accounting for syntactic complexity, semantic composition, surprisal or referent order. We interpret this as evidence of the behavioral cost of two discourse-level mental operators: +simple—i.e., adding a simple referent to the mental model, or +composite—building a composite referent from simpler ones. Whether the two operators are cognitively distinct remains inconclusive. These results shed light on an important building block in discourse-level comprehension, and spawn hypotheses about hierarchically structured discourse representations during language comprehension.

References

Agmon, G., Pradhan, S., Ash, S., Nevler, N., Liberman, M., Grossman, M., & Cho, S. (2024). Automated Measures of

Syntactic Complexity in Natural Speech Production: Older and Younger Adults as a Case Study. *Journal of Speech, Language, and Hearing Research*, *67*(2), 545–561. https://doi.org/10.1044/2023_JSLHR-23-00009

Brown-Schmidt, S., Byron, D. K., & Tanenhaus, M. K. (2005). Beyond salience: Interpretation of personal and demonstrative pronouns☆. *Journal of Memory and Language*, *53*(2), 292–313. https://doi.org/10.1016/j.jml.2005.03.003

Coopmans, C. W., & Nieuwland, M. S. (2020). Dissociating activation and integration of discourse referents: Evidence from ERPs and oscillations. *Cortex*, *126*, 83–106. https://doi.org/10.1016/j.cortex.2019.12.028

Duff, J., & Altshuler, D. (2025). *Reanalysis in discourse comprehension: Evidence from reading times*. 102–110.

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

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language*, *26*(1), 69–83. https://doi.org/10.1016/0749-596X(87)90063-5

Haitao Liu. (2008). Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, *9*(2), 159–191. https://doi.org/10.17791/JCS.2008.9.2.159

Johnson-Laird, P. N., & Garnham, A. (1980). Descriptions and Discourse Models. *Linguistics and Philosophy*, *3*(3), 371–393. JSTOR.

Lakens, D. (2017). Equivalence Tests: A Practical Primer for *t* Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Murphy, G. L. (1984). Establishing and accessing referents in discourse. *Memory & Cognition*, *12*(5), 489–497. https://doi.org/10.3758/BF03198311

Nieuwland, M. S., Coopmans, C. W., & Sommers, R. P. (2019). Distinguishing Old From New Referents During Discourse Comprehension: Evidence From ERPs and Oscillations. *Frontiers in Human Neuroscience*, *13*, 398. https://doi.org/10.3389/fnhum.2019.00398

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*.

Rayner, K., Garrod, S., & Perfetti, C. A. (1992). Discourse influences during parsing are delayed. *Cognition*, *45*(2), 109–139. https://doi.org/10.1016/0010-0277(92)90026-E

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1908.10084

Richter, T., Singer, M., Rapp, D. N., Schober, M. F., Britt, M. A., Britt, M. A., Rapp, D. N., & Schober, M. F. (2018). Discourse Updating: Acquiring and Revising Knowledge through Discourse. In *The Routledge Handbook of Discourse Processes* (2nd ed., pp. 167–190). Routledge.

Schuh, J. M., Eigsti, I.-M., & Mirman, D. (2016). Discourse comprehension in autism spectrum disorder: Effects of working memory load and common ground: Working Memory, Common Ground and Discourse in ASD. *Autism Research*, *9*(12), 1340–1352. https://doi.org/10.1002/aur.1632

Tulling, M., Law, R., Cournane, A., & Pylkkänen, L. (2021). Neural Correlates of Modal Displacement and Discourse-Updating under (Un)Certainty. *Eneuro*, *8*(1), ENEURO.0290-20.2020. https://doi.org/10.1523/ENEURO.0290-20.2020

Wittenberg, E., Momma, S., & Kaiser, E. (2021). Demonstratives as bundlers of conceptual structure. *Glossa: A Journal of General Linguistics*, *6*(1). https://doi.org/10.5334/gjgl.917

Yngve, V. H. (1960). A Model and an Hypothesis for Language Structure. *Proceedings of the American Philosophical Society*, *104*(5), 444–466. JSTOR.

Zwaan, R. A., & Madden, C. J. (2004). Updating situation models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 283–288. https://doi.org/10.1037/0278-7393.30.1.283