Title
Discourse structure affects reference resolution to events in English: Evidence from a new paradigm

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Wampler, Joshua
Wittenberg, Eva

Publication Date
2023

Peer reviewed

Discourse structure affects reference resolution to events in English: Evidence from a new paradigm

Joshua Wampler (jwwample@ucsd.edu)
University of California, San Diego, Department of Linguistics, 9500 Gilman Dr,
San Diego, CA 92161, United States

Eva Wittenberg (wittenberge@ceu.edu)
Central European University, Department of Cognitive Science, Quellenstraße 51
Vienna, 1100, Austria

Abstract

Reference to events using pronouns like *it* or demonstratives like *that* has been difficult to study, because unlike in reference to people or objects, the ground truth of interpretation is harder to establish. In this paper, we introduce a new task to understand the roles of three parameters in event reference resolution: The referential expressions themselves, sentential aspect, and discourse structure. We find that different referring expressions reliably refer to specific parts of a discourse, and confirm previous findings that sentential aspect influences referential accessibility; that the structure of a discourse itself has a major effect on reference resolution, and most notably that, contrary to predictions in the literature, some events that are not on the right-frontier of a discourse are nevertheless available for reference. Our findings contribute to a growing literature on anaphor resolution that finds that the parameter structure that determines reference resolution is multifaceted, but predictable.

Keywords: event reference; Gantt charts; pronouns; proforms; demonstratives; anaphor; discourse structure

Introduction

In every conversation, people talk about and refer to events, using pronouns like *it*, demonstratives like *that*, and adverbial forms like *so* and *thus,* as exemplified in (1):

(1) Dora and Mona went to the CogSci conference 2022, which happened in Toronto. *It* was extremely interesting and lasted three days. *That* was just the right amount of time. They listened to each other's talks, and doing *so* required quite some mental effort, although *it* is easier than on zoom. Mona, like Dora, hopes to go back to in-person conferences. She also wants to keep hybrid models. Doing *thus* facilitates access for those who cannot travel.

Reading a text like this, it feels effortless to connect the referring expressions in italics to the events they refer to: *It* refers to the conference, *that,* to the duration, *so*, to the action of listening, and so on.[1] But what are the defining parameters that establish the link between a referring expression and its potential eventive referent?

In this paper, we manipulate three parameters to better understand event reference resolution: The referential expressions themselves, sentential aspect, and discourse structure. These factors have been found to be crucial in personal pronoun resolution, but it is unclear how and whether findings on personal pronouns like *she* or *he*, which in English refer to people or other animate entities, translate to event reference.

We use proforms like *it* and *so*, and demonstratives like *that* or *thus*, following recent studies showing that form-specific preferences are strongly predictable of the resolution of object and event reference. In English, demonstratives such as *that* tend to refer to events and other complex cognitive construals more than simple pronouns such as *it* (Bevacqua et al., 2021; Brown-Schmidt et al. 2005; Çokal et al. 2014, 2018; Marx et al., under review; Wittenberg et al., 2021). However, in these previous studies, the potential referents for *it* and *that* always included at least one object that *it* could potentially refer to, such as *lasagna* in (2):

(2) Adam made lasagna for me last night.
  a. It was amazing.
  b. That was amazing.

Within the domain of event reference, it is therefore an open question how proforms like *it* or *so,* or demonstratives such as *that* and *thus,* are linked to their eventive referents.

The second factor we study is sentential aspect. Imperfective and perfective aspect have different properties; most relevant for this study, in English, imperfective is claimed to present events as "open", while perfective aspect presents events as "closed". In consequence, imperfective aspect facilitates access to the internal participants of an event, while perfective focuses attention on the result state, or the event as a whole (Bevacqua et al., 2021; Ferretti et al., 2009; Wampler, 2021). If this is true, then subevents of an event encoded with imperfective aspect should be preferred for reference over subevents of an event encoded with perfective aspect.

Finally, we also investigate the role of discourse structure, based on Webber's (1991) claims that only referents corresponding to discourse nodes on the right frontier of the structure are sufficiently accessible for reference. However, from work on personal pronouns, we know that semantic enrichment through elaboration on a referent can facilitate

[1] Following Wittenberg et al. (2021), we use 'referent' to denote the linguistic and conceptual entity a proform or demonstrative refers to, and 'event' to generally mean 'things that happen over time' (Casati & Varzi, 2008).

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

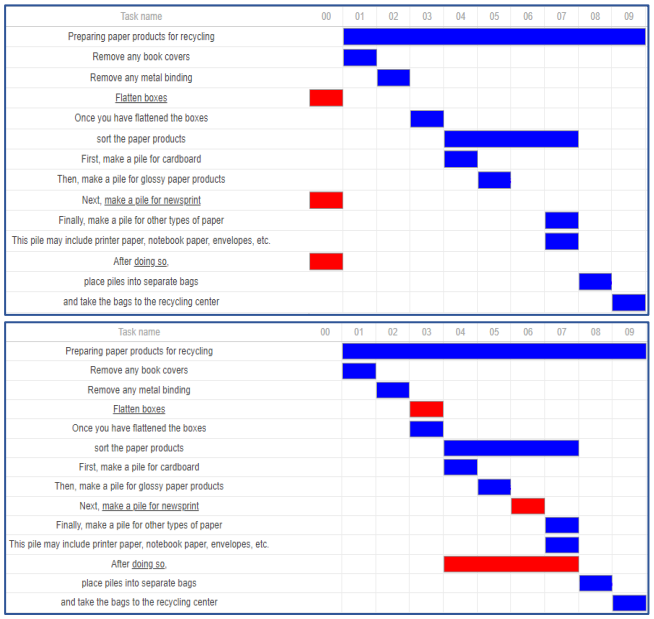

Figure 1. Illustration of the Gantt Chart paradigm, before (top) and after being filled out (bottom).

retrieval, possibly due to increased activation (Hofmeister, 2011; Troyer et al., 2016, Karimi et al., 2020). That is, while the right-frontier theory would predict that the whole discourse and the last event in a discourse should be privileged for event anaphora, this semantic enrichment account would predict that other events that contain complex descriptions of subevents should be more accessible over simple events.

Event reference has been studied from multiple angles in linguistic theory, computational linguistics, and sentence processing. However, understanding the factors influencing event reference has proven difficult for several reasons, one of which is the uncertainty tied to almost any eventive anaphora. Consider reference to person as comparison: The pronouns *he*, *she*, and singular *they* can be ambiguous in any context, but they are never ambiguous across conceptual categories: You may not be sure whether *she* in (1) refers to Dora or Mona, but you can be reasonably sure that *she* refers to one of the people mentioned in the discourse before. In contrast, the referent of an eventive or object anaphor may not always (or even often) have a neat linguistic precursor: For the second *it* in (1), there are numerous referential candidates, from *mental effort* to the – in itself anaphoric – *doing so;* another candidate is the whole previous discourse, or *it* may be expletive. This rich, less restricted space of referential possibilities poses problems for both computational (Kolhatkar et al., 2018; Poesio et al., 2023) and psycholinguistic approaches (e.g., Wittenberg et al., 2021; but see Bevacqua et al., 2021).

Here, we overcame these problems by developing a novel paradigm, the Gantt chart paradigm, to obtain an explicit measure of reference resolution to events (Fig. 1). We report results from a set of online experiments that investigated reference resolution when multiple events are available in the discourse, studying event reference with nominal referring expressions (*it* and *that*; Experiment 1) and adverbial referring expressions (*so* and *thus*; Experiment 2).[2]

Current Experiments

We report two studies that ask 1) whether the type of referring expressions, 2) the aspect of the phrase containing the referring expression, or 3) the structure of the discourse has an effect on which portion of the event structure is chosen as referent.[3] To answer these questions, we developed a novel paradigm, designed specifically for the study of reference to events: the Gantt chart paradigm.

A Gantt chart is a classic visualization technique, illustrating the schedule of a complex event or project as well as dependency relationships between events (Geraldi & Lechner, 2012). Using step-by-step instructions, we utilize a common way to present participants with both a macro-event as well as a number of subevents of which it is composed. Crucially, step-by-step instructions are common in everyday life, and most people can be assumed to be familiar with their structure: An overarching goal, the macro-event, is realized by following a chain of smaller goals, the subevents, which in themselves sometimes contain subevents.

In this study, we used two types of instructions: linear instructions, in which each step directly follows the preceding step, and hierarchical instructions, in which some subevents themselves contain subevents, leading to a nested structure. An example of this can be seen in Figure 1, where the fifth subevent, "sort the paper products", is itself composed of three subevents: making a pile for cardboard, making a pile for glossy paper, and making a pile for newsprint. In the hierarchical stimuli, there was always at least one non-elaborated subevent directly preceding the step with the anaphor (Fig.1). This serves to differentiate predictions made by the right-frontier constraint and the semantic enrichment theory.

In this paradigm, every clause in an instruction set represents one event (step) on its own line in a Gantt chart, corresponding to the time course of the events given by the instructions: The event described in Step 1 is followed by that described in Step 2, which is followed by that described in Step 3, and so on (Fig. 1, top). The Gantt chart is largely pre-filled with blue cells, indicating when each step occurs in relation to other steps. For some steps (critical and filler steps), no cells are filled in. These steps are underlined and correspond to a red cell. Participants place the red cell in the appropriate location on the Gantt chart (Fig. 1, bottom). The red cells can be moved to the left or the right, and be

[2]Both experiments were replicated in the lab, producing similar patterns of results to those reported here; data and analyses can be found on https://osf.io/347zs/.

[3]General preregistration for both studies (and their in-lab replications) under https://osf.io/6cm95; code and data under https://osf.io/347zs/. A video demonstrating the use of this paradigm can be found at https://osf.io/q2ujn.

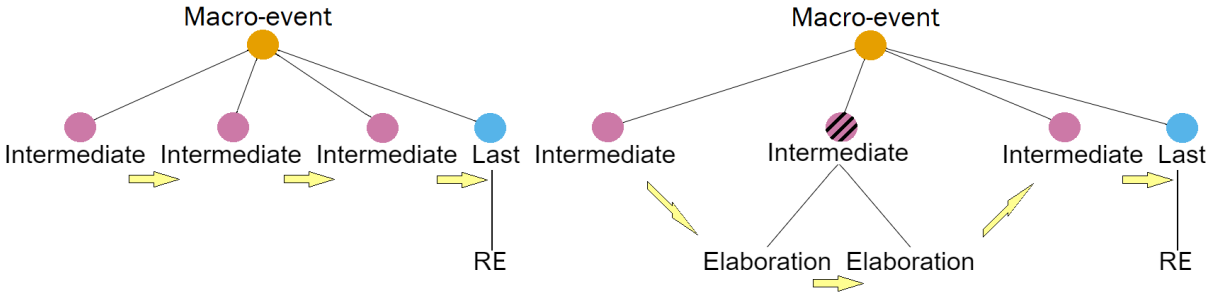

Figure 2: Types of event structures: Linear events (left) in which the macro-event contains a linear string of intermediate subevents, until the last event is reached; and hierarchical events (right), in which an intermediate event contains one or more elaborations. Yellow arrows indicate the temporal ordering of the subevents.

expanded or narrowed. Importantly for our experiments, these red cells can be taken as the product of a reference resolution process.

This experimental paradigm has several advantages: As opposed to indirect measures such as reaction times, it allows us to probe participants' *interpretation* of referring expressions; it is sufficiently intuitive to be learned quickly; and the dependent variable, proportion of possible cells selected, was quantifiable and generalizable over different event-structural manipulations. One disadvantage of the paradigm is its somewhat metalinguistic nature; however we believe this would be unavoidable for any workable paradigm, given the unique experimental constraints involved in studying event anaphora.

Predictions

Predictions about the effect of referring expression: First, in analogy with findings from previous literature (Bevacqua et al., 2021; Brown-Schmidt et al. 2005; Çokal et al. 2018; Wittenberg et al., 2021; Marx et al., under review), both the type of referring expression (proform vs. demonstrative) should affect the complexity of the chosen referent: Proforms should be resolved to less complex referents than demonstratives.

Predictions about the effect of aspect: On analogy with reference to people (Ferretti et al., 2009), we predict that imperfective will lead to more resolution to intermediate steps (purple in Fig. 2), while perfective will lead to more resolution to the macro-event (yellow) or the last step in the preceding discourse (blue).

Predictions about the effect of event structure: We predict discourse structure to have an effect on which portion of the discourse is chosen as referent. If intermediate steps (purple in Fig. 2) with nested elaborations (hierarchical) are more salient for reference than events without such elaboration (linear), then we predict that hierarchical discourse will lead to more intermediate events being considered as referents than in linear discourse.

Specific predictions for hierarchical events: Two predictions arise for hierarchical events only, regarding the interaction of aspect and structure: First, in hierarchical discourse, referring expressions under imperfective aspect should be resolved to referents corresponding to elaborated subevents in the structure more often than referring expressions under perfective aspect. Second, referring expressions embedded under perfective aspect will be resolved to referents corresponding to the macro-event in the structure more often than under imperfective aspect.

Experiment 1

Participants: We recruited 160 self-declared native speakers of English via Amazon Mechanical Turk. We used CloudResearch (Litman, Robinson, & Abberbock, 2017) services, restricting the participant pool to users with an IP address in the United States, a completed task acceptance rate of 80% or higher, and at least 100 tasks completed. Participants were compensated at an average rate of $12/hr.

Materials and Procedure: Twelve stimuli sets were created from modified instructions found on the internet. We manipulated aspect (perfective: *done that*, vs. imperfective: *doing that*), and referring expression (*it* vs. *that*) within subjects. The dependent variable was the proportion of cells selected for the step containing the referring expression.

Additionally, we manipulated the structure of the instructions, such that half had a linear and half a hierarchical structure. These different structures were crossed with aspect, such that there were three perfective and three imperfective linear and hierarchical structures, respectively.

Instruction sets ranged in length from 13 to 23 steps (average: 17) in order to prevent repetition effects. Each instruction set had one moveable red cell associated with the step containing the referring expression. Additionally, each instruction set contained several other moveable red cells as distractors, the number of which varied depending on the instruction length (≤16 steps = 3 distractors; >16 steps = 4 distractors; average distractors per stimuli: 3.75). In addition to the twelve critical instruction sets, 24 fillers were created. The fillers did not include any of the referring expressions under investigation.

All manipulations were run within-subjects. Two different lists in a pseudorandomized order were created, such that there were always two fillers between critical instruction sets. Participants were randomly assigned to a list using a Latin square design.

Analyses: For both experiments, the dependent variable was the proportion of possible cells selected. Cells selected were further coded for type: macro-event, (elaborated)

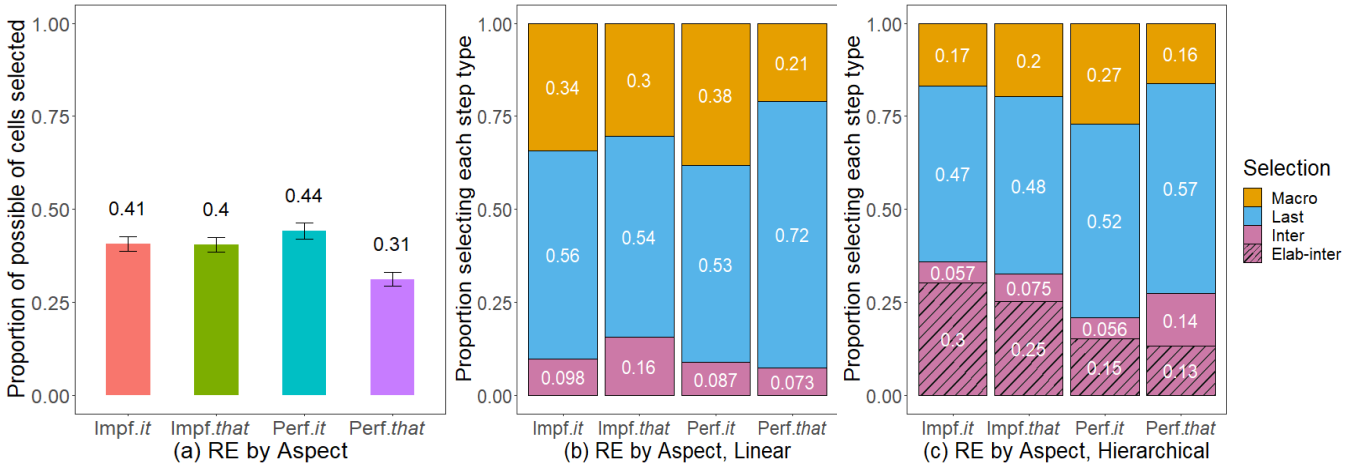

Figure 3: Results of Experiment 1: (a) Average proportion of cells selected, by referring expression (RE; *it*, *that*) and aspect (imperfective, perfective) , error bars indicate standard error; (b) Average proportion of total selections for each step type in linearly structured stimuli, by referring expressions (RE; *it*, *that*) and aspect (imperfective, perfective); (c) Average proportion of total selections for each step type in hierarchical discourse, by referring expressions (RE; *it*, *that*) and aspect (imperfective, perfective).

intermediate step, or last step before the referring expression. We used proportion of possible cells, as opposed to absolute number of cells, because the number of cells preceding the referring expressions was intentionally variable.

Two linear mixed-effects models and one multinomial model were constructed, which all included subject and item as random effects. Independent factors were contrast-coded. The main model tested our first two predictions, that there should be an effect of referring expression or aspect. It contained referring expression (*it*/*that*) and aspect (imperfective/perfective) as fixed effects. The second model was built to test whether different event types were selected depending on structure. It contained the selected event type ((elaborated) intermediate step or not) as the dependent variable, and structure (linear/hierarchical) as fixed effect.

Finally, a multinomial logistic model tested predictions that were specific to the hierarchical instructions. This last model contained the selected event type (macro-event/ intermediate step/elaborated intermediate step/last step) as dependent variable, and aspect (imperfective/perfective) as fixed effect. The reference category for type of step was "last step"; reference for aspect was perfective.

Results: Participants selected more complex referents for perfective *it*, and less complex referents for perfective *that*, than for imperfective *it* or *that*, respectively (Fig. 3a). Both main effects of referring expression ($p < 0.001$; $\beta = 0.03$; $t = 4.8$) and aspect ($p < 0.01$; $\beta = 0.01$; $t = 0.8$), as well as their interaction ($p < .01$; $\beta = -0.02$; $t = -3.4$) were significant.

Table 1: Summary statistics for multinomial model, Experiment 1

		(Intercept)	Perfective
Coefs.	Macro-event	-0.957	-0.014
	Elaborated intermediate	-0.937	0.412
	Intermediate	-1.83	-0.16
Std. Err.	Macro-event	0.098	0.098
	Elaborated intermediate	0.099	0.099
	Intermediate	0.141	0.141

The structure of the discourse also affected which steps were selected for reference, with participants selecting (elaborated) intermediate steps more often in hierarchical discourse than in linear discourse (main effect of structure: $p < 0.001$; $\beta = -0.83$; $z = -3.9$; Figs. 3b,c).

Within hierarchical instructions (Table 1), participants selected elaborated intermediate steps more often when the referring expression was embedded under imperfective aspect ($\bar{x} = 0.28$, $sd = 0.45$) than under perfective aspect ($\bar{x} = 0.14$, $sd = 0.35$). Conversely, participants selected macro-events more often when the referring expression was embedded under perfective aspect ($\bar{x} = 0.21$, $sd = 0.41$) than under imperfective aspect ($\bar{x} = 0.18$, $sd = 0.39$).

Discussion: We predicted that both **referring expression** and **aspect** would influence reference resolution, with *it* under imperfective aspect leading participants to select less complex referents than *that* under perfective aspect. Though we found both an effect of referring expression and aspect, the pattern was more complex, and not as predicted: The form of the referring expression appears to have no effect when it is embedded under imperfective aspect, while under perfective aspect, the results ran counter to our predictions: perfective *it* resolved to more complex referents than perfective *that*, and it was also resolved to more complex referents than imperfective *it* (and *that*). In turn, imperfective aspect led to less complex referents (as predicted) than perfective aspect under *it*, but more complex referents under *that*. Thus, our predictions were only partially correct, which we will examine in more detail in the General Discussion.

Furthermore, we confirmed our predicted effect of **structure**, with hierarchically structured discourse leading participants to select more (elaborated) intermediate steps than in linearly structured discourse. It is worth noting that the proportion of non-elaborated intermediate steps was relatively stable between both hierarchical and linear discourse; the difference was driven by the fact that the elaborated intermediate steps in particular were often chosen

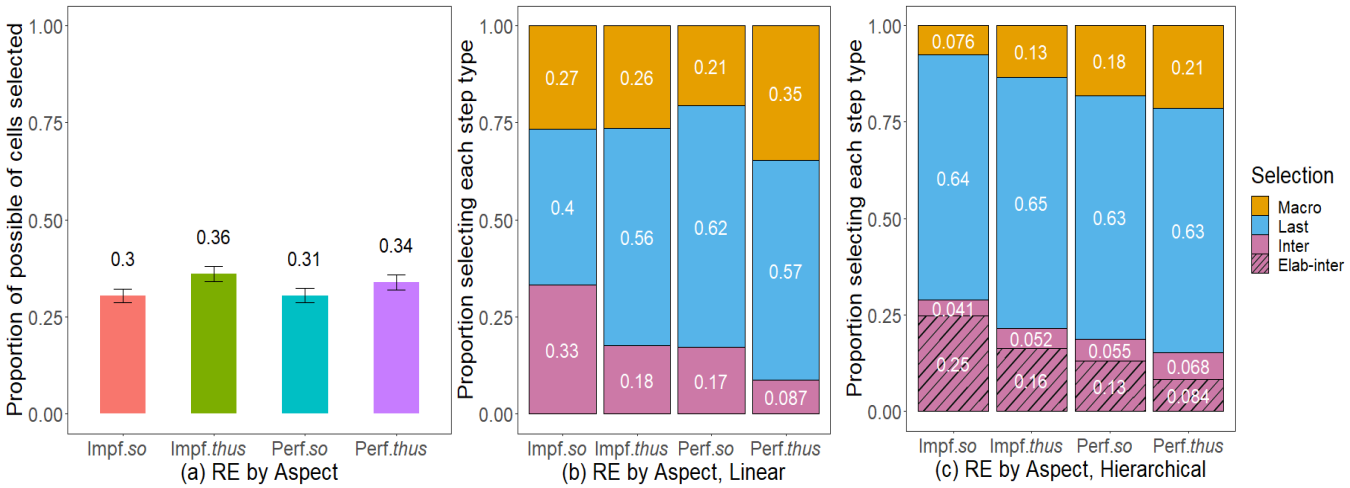

Figure 4: Results of Experiment 2: (a) Average proportion of possible cells selected, by referring expression (RE; *so*, *thus*) and aspect (imperfective, perfective), error bars indicate standard error; (b) Average proportion of total selections for each step type in linearly structured stimuli, by referring expression (RE; *so*, *thus*) and aspect (imperfective, perfective); (c) Average proportion of total selections for each step type in hierarchical discourse, by referring expression (RE; *so*, *thus*) and aspect (imperfective, perfective).

as referents, and that such events are by definition only available in hierarchical discourse.

Within the hierarchically structured stimuli, we predicted that there would additionally be a two-fold effect of aspect, with imperfective aspect leading participants to select more elaborated intermediate steps than perfective aspect, while perfective aspect would lead participants to select more macro-events than imperfective aspect. This prediction was confirmed as well: Participants selected elaborated intermediate steps for reference more often when the referring expression was embedded under imperfective aspect than under perfective aspect, and macro-events more often under perfective aspect than under imperfective aspect. The full picture is slightly more complicated, however. Splitting the data by referring expression as well as aspect reveals that, while perfective *it* led to more selections of the macro-event than imperfective *it*, perfective *that* led to less selections of the macro-event than imperfective *that*.

This can be explained by noticing that, though perfective *that* led to less selection of the macro-event, it also led to more selections of the last event. As noted in the introduction, perfective aspect tends to focus attention on either the event as a whole, or the result state. It appears that these possibilities are split between *it* and *that*, with perfective *it* preferentially being interpreted as referencing the event as a whole (the macro-event), and perfective *that* as referencing the end result (the last event).

So far, we have investigated how people resolve event reference by nominal referring expressions such as *it* and *that*. With object reference, the referring expression must be such a nominal, but when referencing events, the referring expression may instead be an adverbial, such as *so* or *thus*. On the one hand, since *so* and *thus* have been suggested to be adverbial analogs of *it* and *that* (Wampler, 2021), we might expect reference with adverbials and nominals to behave similarly. On the other hand, this may not be the case, as analogy between the two types of expressions is not perfect. Both *it* and *that* may refer to either objects or events, while *so* and *thus* may only refer to events; there is no similar pair that can only refer to objects. Experiment 2 therefore asks whether *so* and *thus* will pattern similarly to *it* and *that* in the domain of event reference.

Experiment 2

Participants: 160 participants were recruited for this study through Amazon Mechanical Turk according to the same criteria as in Experiment 1, and gave informed consent. Participants were compensated at an average rate of $12/hr.

Materials and Procedure followed Experiment 1, with *so* and *thus* replacing *it* and *that*, respectively.

Analyses and Results: The models constructed were the same as in Experiment 1. We found a main effect of referring expression ($p < 0.01$; $\beta = -0.02$; $t = -3.0$), but no significant effect of aspect ($p = 0.5$; $\beta = 0.009$; $t = 0.68$), and no significant interaction ($p = 0.53$; $\beta = -0.004$; $t = -0.63$). These results show that the form of the referring expression significantly affected chosen referent complexity, with participants selecting more complex referents for *thus* ($\bar{x} = 0.35$, $sd = 0.37$) than for *so* ($\bar{x} = 0.31$, $sd = 0.34$; Fig. 4a).

Turning to the question of how discourse structure affects referent selection, we found a main effect of structure ($p < 0.05$; $\beta = -0.57$; $z = -2.2$) in the multinomial model, showing that that participants selected (elaborated) intermediate steps for reference more often in hierarchically structured ($\bar{x} = 0.21$, $sd = 0.41$) discourse than in linearly structured discourse ($\bar{x} = 0.11$, $sd = 0.31$; Figs. 4b,c).

There was also a main effect of aspect ($p < 0.001$) within hierarchically structured instructions (Table 2): Participants selected intermediate steps for reference more often when the referring expression was embedded under imperfective aspect ($\bar{x} = 0.21$, $sd = 0.4$) than under perfective aspect ($\bar{x} = 0.11$, $sd = 0.31$). Additionally, the macro-event was selected more often for referring expressions embedded under

perfective aspect ($\bar{x}$ = 0.2, *sd* = 0.4) than under imperfective aspect ($\bar{x}$ = 0.11, *sd* = 0.31).

Table 2: Summary statistics for multinomial model, Experiment 2

		(Intercept)	Perfective
Coefs.	Macro-event	-1.485.	-0.014
	Elaborated intermediate	-1.46	-0.325
	Intermediate	-2.48	-0.146
Std. Err.	Macro-event	0.112	0.112
	Elaborated intermediate	0.11	0.11
	Intermediate	0.169	0.169

Discussion: We had predicted that both referring expression and aspect would influence reference resolution. We found only an effect of referring expression and no effect of aspect: In general, *so* was resolved to less complex referents than *thus*, as predicted. This pattern held when the conditions were crossed, with imperfective *so* resolving to less complex referents than imperfective *thus*, and perfective *so* resolving to less complex referents than perfective *thus*; the differences between aspects were not significant.

We additionally confirmed a predicted effect of structure: People selected proportionally more (elaborated) intermediate steps in hierarchical discourse, than in linear discourse. This was true for every condition except imperfective *so*, which lead to selection of more (elaborated) intermediate steps in linear than in hierarchical discourse.

Within the hierarchically structured stimuli, we also confirmed our prediction that there would additionally be a two-fold effect of aspect, with imperfective aspect leading participants to select more elaborated subevents than perfective aspect, while perfective aspect would lead participants to select more macro-events than imperfective aspect. The more complicated split observed in Experiment 1 for perfective *that* did not reappear here, for either *thus* or *so*.

General Discussion

Our results suggest that both event structure and aspect play a role in resolving event reference: Elaboration of an event raises its salience for reference, and imperfective aspect leads to more resolution to (elaborated) subevents, while perfective leads to more resolution to the macro-event and the last subevent. These results support findings that elaborations can raise activation levels in working memory and extend work on the interaction of aspect and reference resolution. Aspect not only affects access to the internal participants of simple events, but also to the internal subevents of complex events.

Further, these results call into question the strength of the right-frontier constraint. Webber (1991) claims that only the right-frontier of a discourse structure is available for reference, as it is only this portion of the discourse that is in focus, i.e., salient enough for reference. The constraint makes no allowance for salience derived from sources other than the right-frontier. Contrary to the right-frontier theory, our results suggest that elaborating information can raise the salience of a non-right-frontier discourse segment enough to overcome any tendency towards the right-frontier. The constraint would need to be reframed as a calculation of salience that takes into account not only position in the discourse structure, but other sources of salience as well.

We predicted that proforms would be resolved to less complex referents than demonstratives, but only adverbial referring expressions resulted in this pattern. With nominal referring expressions, we found a complex interaction between the form of the referring expression and the aspect under which it was embedded.

So, the question is, why was *it* not resolved to less complex referents than *that*? A possible explanation may lie in the type of search instructions encoded in different referring expressions. It is thought that demonstratives lead hearers to search for a more complex referent, perhaps having to create a composite referent on the fly, while proforms like *it* lead hearers to search for a simple, ready-to-use referent (Brown-Schmidt et al., 2005; Marx et al., 2023; Wittenberg et al., 2021). In our stimuli, the individual steps were clearly parts of the larger macro-event. As such, a simple, readily available referent may have been the macro-event, and separating any of the subevents from the macro-event may have been a more complicated procedure. If this is the case, then it is not surprising that in the perfective, which biases reference towards the macro-event or the last subevent, *it* resulted in more cells being selected since participants selected the macro-event more often, and *that* resulted in fewer, since participants selected the last subevent only.

Another question is why we observed slightly different results between nominal (Experiment 1) and adverbial (Experiment 2) referring expressions. Adverbial referring expressions unambiguously target event referents, while nominal referring expressions may target many types of referents. Even embedding a nominal referring expression under the verb *do* does not entirely clear up this ambiguity, as in *Look at my new painting. I did it yesterday*, where *it* refers to the new painting. As such, it is reasonable to assume that the algorithm used in resolving reference for nominal referring expressions must consider more factors than that used for adverbial referring expressions. It may be this additional complexity that results in the observed split, and that we simply do not yet have the full picture of what goes into the computation of reference. Regardless, it is clear that the nominal/adverbial analogy proposed in Wampler (2021) does not explain our data in full.

Finally, we want to draw attention to the effectiveness of our experimental design. Studying event reference has been notoriously difficult in experimental settings; events are harder to individuate than objects, often falling into part-whole relationships with other events. By utilizing Gantt charts to represent complex events composed of smaller subevents, we were able to visually represent these mereological aspects of events in an intuitive way. With minimal training, participants were able to consistently use Gantt charts to clearly indicate which portion of the event structure they took a referring expression to be referring to. It is our hope that others will take up this paradigm as a useful tool for further exploration of the domain of event reference, as well as discourse structural questions more generally.

References

Bevacqua, L., Loáiciga, S., Rohde, H. and Hardmeier, C., 2021. Event and entity coreference across five languages: Effects of context and referring expression. *Dialogue & Discourse*, *12*(2), pp.192-226.

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.

Casati, Roberto & Achille C. Varzi (2008). Event concepts. In Thomas F. Shipley & Jeffrey M. Zacks (eds.), *Understanding Events: From Perception to Action*, 31–53. Oxford: Oxford University Press.

Çokal, D., Sturt, P., & Ferreira, F. (2014). Deixis: *This* and *that* in written narrative discourse. *Discourse Processes*, *51*(3), 201-229.

Çokal, D., Sturt, P., & Ferreira, F. (2018). Processing of *it* and *this* in written narrative discourse. *Discourse Processes*, *55*(3), 272-289.

Ferretti, T. R., Rohde, H., Kehler, A., & Crutchley, M. (2009). Verb aspect, event structure, and coreferential processing. *Journal of Memory and Language*, *61*(2), 191-205.

Geraldi, J., & Lechter, T. (2012). Gantt charts revisited: A critical analysis of its roots and implications to the management of projects today. *International Journal of Managing Projects in Business*, 5.4, pp. 578- 594.

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Language and cognitive processes*, *26*(3), 376-405.

Karimi, H., Diaz, M. T., & Wittenberg, E. (2020). Sheer Time Spent Expecting or Maintaining a Representation Facilitates Subsequent Retrieval during Sentence Processing. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society,* pp. 2728-2734.

Kolhatkar, V., Roussel, A., Dipper, S., & Zinsmeister, H. (2018). Anaphora with non-nominal antecedents in computational linguistics: A survey. *Computational Linguistics*, *44*(3), 547-612.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2), 433-442.

Marx, E., Iwan, O. & author 2 (under review). When you say that, do you mean it? Crosslinguistic patterns of anaphor resolution in English, German, and Polish.

Poesio, M., Yu, J., Paun, S., Aloraini, A., Lu, P., Haber, J., & Cokal, D. (2023). Computational Models of Anaphora. *Annual Review of Linguistics*, *9*.

Troyer, M., Hofmeister, P., & Kutas, M. (2016). Elaboration over a discourse facilitates retrieval in sentence processing. *Frontiers in Psychology*, *7*, 374.

Wampler, J. (2021). Do thus: an investigation into anaphoric event reference. *Glossa: A Journal of General Linguistics*, *6*(1).

Webber, B. L. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, *6*(2), 107-135.

Wittenberg, E., Momma, S., & Kaiser, E. (2021). Demonstratives as bundlers of conceptual structure. *Glossa: A Journal of General Linguistics*, *6*(1).

Acknowledgments

The authors would like to thank the amazing undergraduate research assistants at the former Language Comprehension Lab at UC San Diego, without whom there would be no experiments to report: Mohit Gurumukhani, Nico Hammer, Alex Jung, Shubam Kausal, Harrison Kim, Zhiyi Li, Jessica Luo, Allison Park, Elizaveta Pertseva, Ruoqi Wei, Emma Wilkinson, Jane Yang, Lea Zaric, and Alexandra Zenteno; and the members of UC San Diego's Semantics Babble discussion group, for enlightening discussion and critiques of this work throughout its development.