Creating Coherence with Concepts

Mandy Simons, Carnegie Mellon University
David Danks, University of California, San Diego
October 2021

Abstract

Where does coherence reside? In current treatments of discourse coherence, coherence is understood as a property of a text or discourse, attributable on the basis of relations that hold between its proposition-denoting units. In this paper, we offer an alternative (but compatible) cognitively-based view of coherence, which locates coherence in the agent's mental representation of the content of the discourse, and in the processes whereby this representation is generated. On this view, coherence reflects the presence of connections between the *concepts* which are activated in the course of processing. These conceptual interelations are not necessarily (or typically) propositional. In the paper, we lay out this notion of *concept-level coherence*, and illustrate it with an analysis of nominal bridging.

1. What is coherence a property of?

Within linguistics, coherence is taken to be a property of a discourse, a linguistic object consisting of a series of connected utterances. The two main contenders for a theory of discourse coherence are SDRT (Segmented Discourse Representation Theory: Asher & Lascarides 2003 i.a.), and the coherence relations theory of Kehler (2002, 2012), which itself draws on prior work by Hobbs (1979, 1990) . The central notion in both SDRT and in Kehler's work is that of a coherence relation, a relation which holds between units of discourse (typically, interpreted clauses) and characterizes how the units are related to each other. In addition, SDRT offers a characterization of discourse structure. Structure arises from the relations built between elementary discourse units, creating complex discourse units. SDRT posits a central structural constraint, the Right Frontier Constraint, which limits the way in which a new discourse unit can connect to the ongoing discourse, given its rhetorical structure. In SDRT, a discourse is coherent just in case a structure built of these relations can be constructed.

An alternative, not mutually exclusive, understanding of coherence locates this feature not in the linguistic material, or even in an abstract representation of content, but in the mental representations or world model that the hearer constructs as they process and comprehend extended linguistic input. On this view, a hearer judges a discourse to be coherent when the mental

representation that is triggered by the linguistic input is coherent (in a sense to be clarified in the course of this paper). Coherence of a discourse is, on this view, based only *indirectly* on surface features of the discourse itself, inasmuch as those features lead to coherent mental representations. Certainly, actual judgments of discourse coherence may reflect properties of *both* linguistic form and mental representation; ordinary judgments of coherence are undoubtedly sensitive to features of the discourse itself, including ordering, appropriate use of signals of anaphoricity and of explicit markers of interclausal connections (words like *because* and *so*), and so on. In this paper, however, we focus on developing an understanding of the role of the hearer's internal representations in creating coherence. Our goal is to add to the linguistic literature on coherence a perspective that so far is absent.

We start from a view of knowledge representation in which concepts are the central building block. Concepts are representations of our knowledge and beliefs about kinds of things and events, including the myriad relations between them. We propose that coherence of mental representations involves integration of the concept-based representations generated during comprehension. We provide a more specific account in section 2, but begin here with a preliminary sketch of the overall picture.

In modeling reasoning with concepts, it is important to distinguish full conceptual representations (held in long term memory) from instantiations of concepts, where the latter are elements in working memory that represent particular instances of a concept, and that are actively used in a mental process – such as language interpretation. For example, there is a difference between one's concept of DOG, and one's representation of Fido. We make the simple assumption that one aspect of language processing involves the introduction into working memory of instantiations of the concepts which the speaker invokes in their utterances. For example, if a speaker says "I have a dog," then the hearer will (amongst other processes) instantiate their concept DOG in working memory in order to represent the specific dog under discussion.

When a concept is initially instantiated, many of its features will remain indeterminate. When the speaker says "I have a dog," the hearer will instantiate the concept DOG, but cannot fix values for any basic features such as size, breed, color and so on. Now suppose that the speaker continues: "It's a dachshund." The speaker is now able to assign a value to one of these features, elaborating the instantiation that constitutes their current representation of the discourse content. In our model,

this is how the second utterance coheres with the first: by virtue of the fact that it can be used to elaborate the instantiation "launched" by the first.

Now suppose that the speaker's next utterance is: "It's a greyhound." At this point the hearer is likely to be confused, because a dog can't simultaneously be a dachshund and a greyhound. At least informally, we would be inclined to say that the sequence of utterances is no longer coherent (unless the new utterance is construed as a correction). In our approach, this judgment of incoherence is a direct reflection of the hearer's world knowledge. Instantiations are constrained by the structure of concepts in long term memory, which directly represent world knowledge about those concepts. For most people, the DOG concept allows only one value for the "breed" variable (with "mixed" as an option). We assume that hearing the speaker say of their own dog, "It's a dachshund" leads the hearer to assign a very high probability to the value "dachshund" for the "breed" variable in their current instantiation of DOG. Hence when the speaker appears to try to attribute a different value to the variable, the instantiation "crashes" -- it cannot be updated to reflect the new information.

In this essay, we develop an account of this type of coherence, which we call *concept-level coherence*. Concept-level coherence contrasts with theories of (what we will call) *proposition-level coherence* that consider coherence only as a relation between (the representations of) clausal contents. The general picture is this: Part of the process of language interpretation is the construction of a complex instantiation involving multiple concepts; this instantiation represents (possibly subparts of) the content of what the speaker says. Once an instantiation is in place, the hearer will use new information provided, as far as possible, to elaborate the existing instantiation. The more straightforwardly information can be integrated in this way, the more coherent the discourse. When interpretation requires the hearer to utilize weak or complex interconnections between concepts, or to modify their existing conceptual structures in order to create new connections, the discourse is judged less coherent.

In what follows, we first give a more detailed explanation of concept-level coherence (Section 2), and then illustrate it with a case study of nominal bridging (Section 3). The idea that bridging is a coherence-driven phenomenon is developed in Asher & Lascarides 1998. Their treatment, though, assumes that the coherence relations in question are the proposition-level coherence relations of SDRT. Here, we offer a view of bridging as driven by instantiation construction and

conceptual coherence. We conclude in Section 4 with a more general discussion of concept-level coherence.

2. An Account of Concept-level Coherence

There is clear consensus that coherence is not a purely linguistic feature of a discourse, but depends also on the hearer's world knowledge, which determines the inferences that they deem obligatory, permissible, or impermissible. As Asher and Lascarides 1998 observe, modeling reasoning with world knowledge is a significant challenge. To address this challenge, we adopt a well-established computational model of knowledge and reasoning in which concepts are represented as probabilistic graphical models. Other computational theories of concepts and knowledge representation could be used instead; our arguments and observations about concept-level coherence are *not* dependent on this particular computational model (though the details do matter for, e.g., the particular account of bridging that we provide in Section 3.2). The key is that the cognitive model of reasoning with world knowledge should provide (1) the right kinds of structures for identifying relations between entities invoked in a discourse, and thus for the kind of inference required for building coherence at the conceptual level; and (2) a cognitively plausible account of this inference. Many different cognitive architectures other than the one we employ here include these two components (though we find this architecture particularly useful and compelling).

We understand concept-level coherence as something that depends on the hearer's ability to integrate the discourse content, and subsequent inferences, into instantiations (tokens) of relevant concepts (types). Concept-level coherence is thus a graded phenomenon; discourses can be more-or-less coherent, not simply coherent or incoherent. (We return to this point in Section 4.) A full account of concept-level coherence thus requires us to provide a model of concepts and of their instantiations. The cognitive architecture we favor uses probabilistic graphical models to encode semantic content (Danks 2007; Goodman et al. 2015; Gopnik et al. 2004; Griffiths & Tenenbaum 2005; Oppenheimer, Tenenbaum, & Krynski 2013; Rehder 2003a, 2003b; Rehder & Hastie 2004; see also the many references in Danks 2014), as it has been shown to have significant explanatory power in several non-linguistic domains. More specifically, we understand concept-level coherence as involving concepts—both type and token—that are represented as probabilistic

graphical models. The same mathematical and computational framework is used for both the type- and token-level content; the difference lies in what operations can be performed on the representations. To avoid confusion between type- and token-level representations, we will use SMALL CAPS to represent the type-level content, and **boldface** to represent the token-level instantiations of that type-level semantic knowledge. We return to concept-level coherence shortly, but first explain some relevant aspects of this cognitive architecture.

The framework of probabilistic graphical models has emerged from statistics, mathematics, and computer science as a powerful way to represent information-bearing relations, whether causal, taxonomic, information-theoretic, or other. These representations are particularly well-suited to capture much of our conceptual knowledge. A significant part of our understanding of BIRD, for example, is knowledge of which factors are relevant for whether something is a bird, as well as the internal relevance relations between those factors. Moreover, this representational framework is unifying, in the sense that the (mathematical) forms of many standard theories of concepts—exemplar, prototype, causal model—can be represented in the language of probabilistic graphical models (Danks 2014).

At a high level, a probabilistic graphical model consists of two components: (i) a directed (usually acyclic) graph—nodes and (possibly directed) edges between those nodes—over the relevant features or dimensions; and (ii) a joint probability distribution (or density) over those same factors. The graph component represents the qualitative relevance relationships, and the distribution/density represents the quantitative relevance relationships. We bind these components together using two assumptions that ensure that the probabilistic graphical model is internally coherent.[1]

As a concrete example, consider the case of a *causal* probabilistic graphical model (i.e., one in which the edges of the graph correspond to direct causal links) about student behaviors: *Studying* → *Knowledge* → *Test performance*. We have a corresponding quantitative probability distribution, P(*Studying, Knowledge*, *Test performance*), which can be expressed as a product (determined by the binding assumptions) of conditional probabilities: P(*Studying*)P(*Knowledge* | *Studying*)P(*Test performance* | *Knowledge*). This full causal graphical model is a compact representation of a

[1]The two assumptions are Markov and Minimality/Simplicity/Faithfulness. In this paper, we can set aside the precise details of those assumptions.

complex, noisy, indeterministic causal structure that can support inferences ("This student performed well; how much do they likely know?"); practical reasoning ("If I want to get a good grade, how much should I study?"); and much more.

In general, we can identify (many) concepts with particular graphical models, and all of the standard inferential operations using concepts map onto probabilistic updating involving those models. For example, feature inference (e.g., "this is a dog; does it have a tail?") corresponds to inference of the value of a variable in a particular graph. Categorization (e.g., "is this thing a dog or a cat?") corresponds to inference about which graph (one per concept) is most likely to be the correct one for this thing. And so forth for other inferences, as well as concept learning or acquisition.

Throughout this discussion, we have referred to 'features' as components of both concepts and probabilistic graphical models. For example, part of the concept DOG is the feature "barks?" that can take multiple values; in the probabilistic graphical model corresponding to DOG and its instantiations, we therefore have a node (in the graph) and variable (in the probability distribution) labeled "barks?" Importantly, these features can themselves be concepts with their own structure; that is, the graphical model *nodes* can themselves be graphical models that are "encapsulated" in a mathematically precise way so that operations on the larger concept can be independent of the internal structure of each node. We will typically talk about features as primitives, but we do so only provisionally: if necessary, one can use information contained in the concept corresponding to the feature; mathematically, we can "open up" a node to use the graphical model to which it corresponds.[2]

Concepts stand in various relationships to one another. For example, DOG is a (taxonomic) sub-type of ANIMAL, and so we know that any feature of ANIMAL is also a feature of DOG, though perhaps represented only implicitly. There are various complexities around these inter-conceptual relationships; for example, apparent taxonomies might not be strict (Sloman 1998), and the seemingly same concept can appear in multiple types of relationships (Medin et al., 1997). Nonetheless, any framework for representing concepts must allow for these types of relationships. In the case of probabilistic graphical models, this challenge is readily met: since nodes are

[2] There is obviously a potential infinite regress lurking here. There are various ways to either block or embrace that potential regress, each with its own philosophical and psychological advantages and disadvantages. The regress challenge is not, however, relevant for our understanding of coherence.

themselves concepts, we can represent these inter-conceptual relationships as distinctive types of edges in a larger probabilistic graphical model. (See Rubin, Zeigenfuze, & Steyvers 2011 or Danks 2014 for two different ways to model these relations.)

We now return to concept-level coherence. Given this cognitive architecture, we model comprehension as inference and updating over token-level instantiations of concepts, informed by the type-level information in the concepts themselves. During a discourse, an individual has a set of concept instantiations (probabilistic graphical models) encoded in working memory. These instantiations initially encode only the type-level information (as a probabilistic graphical model), and then update feature values as information is acquired. When I hear about my friend's new dog, I have an instantiation **dog** of the concept DOG that encodes epistemic uncertainty about the features of this dog. As I learn more about this dog (e.g., it has three legs), the probability distributions in the instantiation change, while leaving intact the conceptual knowledge represented in DOG.

When the hearer encounters a term,[3] she must incorporate its content into her evolving instantiation. Concept-level coherence reflects her ability to integrate new information into existing instantiations without introducing a large number of new instantiations or significantly overriding previous content (though changes to previous inferences are permitted for dynamic variables). Importantly, the probability of the new information is not particularly important for concept-level coherence. A discourse such as "I have a dog. She has three legs. Her fur is dyed pink." is concept-level coherent, even if the resulting representation is *a priori* quite improbable. In contrast, a discourse such as "I have a dog. I have a car." is not particularly concept-level coherent, even though it describes a much more probable world. The problem with the latter case is that the hearer cannot incorporate the information in the second sentence into a single *integrated* representation, at least not without making a number of additional inferences and introducing a number of new instantiations. The hearer is, on our account, continually updating—through additions and inferences—an integrated world model consisting of concept instantiations, and concept-level coherence tracks the complexity and "effort" of maintaining that integrated representation.

[3] We assume the existence of a process by which hearers connect sound sequences to their corresponding concepts.

Of course, many more details need to be provided about this high-level account. In particular, we need to say what guides the integration of new information into an existing representation. We turn now to the particular case of bridging to focus on this question, and also to illustrate in more concrete detail how concept-level coherence is understood to operate within this cognitive architecture.

3. Bridging: a case study

3.1. Introduction to bridging

A central theme of both SDRT and Kehler's coherence theory is that the expectation (or requirement) that all discourse units can be connected by some coherence relation partly determines the interpretations of the units themselves. To illustrate: A coherent interpretation of example (1) below requires coreference between *she* and *Samara*; given what we know about the transfer of pain, the second clause can be an explanation of the first, as signaled by the presence of *because*, only if the subjects of the two clauses corefer.

(1) Samara is in pain because she stubbed her toe.

Inferered relations between the referents of NPs (Noun Phrases) is not restricted to overtly anaphoric NPs like pronouns. It has long been observed (beginning with Clark 1975) that full NPs are often understood as anaphorically dependent on a prior, non-coreferential NP, resulting in the interpreter inferring some unstated relation between the referents of the NPs.[4] Consider the cases in (2):

(2) Jane looked into one of the rooms.
a. **The ceiling** was very high and
b. **a large window** looked out onto the bay.

The subject NPs in (2)a-b. are interpreted as related to the room that Jane looked into, specifically as parts of that room. However, as Clark also observed, bridges can be much more complex than part-of relations and are highly varied in type. Consider the examples in (3)-(5):

(3) I'm taking my phone back to the store. **The company** has issued a recall.

(4) I like knitted scarves, but **the wool** has to not be itchy.

[4] As Clark noted, bridging relations can also involve events. Here, we focus on cases involving entities introduced by NPs, although in principle the same account is extendable to the event case. See section 4 for some further discussion..

(5) I went to the museum last week. I heard **the admission** was half-off.

As noted above, Asher & Lascarides 1998 propose an account of bridging in terms of proposition-level coherence requirements, arguing that bridged interpretations of NPs arise in order to facilitate the construction of a propositional coherence relation between segments. Asher & Lascarides also observe that world knowledge constrains the construction of bridged interpretations, expressing this in the constraint that "Bridges are plausible", where plausibility reflects "common sense reasoning with world knowledge" (p.97). Indeed, the role of world knowledge in bridging inferences is widely recognized. Clark 1975 notes that bridging inferences "though conveyed by language and a necessary part of the intended message, draw on knowledge of natural objects and events that goes beyond one's knowledge of language itself" (p.412, reprint). For Prince 1992, bridged NPs invoke "Inferrable" entities, where what can be inferred depends on "the hearer's beliefs and reasoning ability." So undoubtedly, a coherence based model of bridging must explain this role for world knowledge.[5]

In SDRT, world knowledge is represented in a set of axioms distinct from the coherence-generating rules; while the two knowledge bases interact, they are independent. In contrast, in the concept-driven approach, world knowledge (as embodied in concepts) is what drives the construction of coherent interpretations. Concept-level coherence emerges from world knowledge, rather than being an independent system, as we illustrated with the dachshund example in Section 1.

In what follows, we will offer some arguments that proposition-level coherence is not adequate as a fully general account of bridging, regardless of one's views about our notion of concept-level coherence. We then briefly discuss the role of definiteness in bridging before turning to the development of our own account.

3.1.1. Bridging and Propositional Coherence

Asher & Lascarides 1998 (see also Hobbs 1979) argue that bridging is a consequence of, and subserves, the construction of coherent discourse structure, and in particular the construction of plausible coherence relations between segments. In many cases (but not universally), bridging will indeed support propositional coherence, as in (6):

[5] See also Bos et al. 1995 and Irmer 2009 for attempts to incorporate aspects of semantic knowledge into an account of bridging within a dynamic framework.

(6) My car isn't drivable. The windshield is cracked.

If the windshield is construed as the windshield of the speaker's car, then the second sentence can be inferred to stand in an Explanation relation to the first. If the windshield in question were any other windshield, it would be hard to identify any coherence relation that holds between the two sentences.

But discourse coherence alone does not seem to provide enough constraints to explain interpretative preferences. For instance, consider the following pair of examples:

(7) I need to fix this old chair. A/the leg is broken. There's no end of things to do.
(8) I need to fix this old chair. A/the leg is broken on that table. There's no end of things to do.

(7) strongly invites a bridged reading of *leg* (whether definite or indefinite) to the chair. Yet there is a plausible and coherent interpretation in which the speaker is referring in (7) to a broken leg of a different item as part of a list of tasks, as shown by (8). If discourse coherence were the only constraint to satisfy, then for (7), both the bridged interpretation (triggering an Elaboration relation) and non-bridged interpretation (triggering a Parallel relation) should be equally accessible.[6] A natural suggestion is that interpreters prefer interpretations that do not require positing new, unmentioned, entities, which suggests that proposition-level coherence is not the sole driver of bridging phenomena. This suggestion follows directly from the account we give below.

A further observation is that a bridging trigger and its anchor may occur within a single Elementary Discourse Unit (or clause). Consider the bridged readings of examples (9) and (10) below.

(9) The school commissioned a teacher to write a report on the issue.
(10) The hospital presented an award to a doctor for excellence in research.

Considerations of discourse coherence do not apply to the interpretation of the NPs *a teacher/a doctor,* again showing that there must be some additional component to the story about coherence.

[6] Asher and Lascarides 1998 do posit partial preference orders on Rhetorical Relations (pp.98-99), but mention only a preference for Explanation over Background. The empirical motivation for this ordering is unclear.

We suggest that what is missing from Asher & Lascarides' account is a consideration of relations that hold between (conceptual representations of) the entities under discussion. The account we offer identifies the role played in bridging by hearers' knowledge of the relations between schools and teachers, hospitals and doctors, or chairs and legs. Certainly, coherence at higher levels of the world model matters as well; interpreters are not only constructing instantiations of entities, but also of events, and these event instantiations must also hold together in a way that is consistent with the world knowledge that the speaker has about event-relations. But in many cases, these entity-level relations suffice to explain bridging, and as noted can explain cases that lie outside the realm of application of proposition-level theories.

3.1.2. Bridging and Definiteness

A different approach to bridging actually does not significantly appeal to coherence at all. The arguably mainstream view found in the literature (see e.g. Clark 1975, Kehler 2015, Roberts 2003) is that *definiteness* plays the central role in triggering bridging: the familiarity implication carried by the definite leads the hearer to search for a way to treat an entity newly introduced into the discourse as "familiar," which can be accomplished by taking it to be related to something previously mentioned. But definiteness is not, in fact, required for bridging. As several researchers have noted (Gundel et al. 1993, Asher and Lascarides 1998, Kehler 2015), indefinites also give rise to bridged interpretations, as we illustrated above in (2)b., (7), (9) and (10). While there are interactions between definiteness and bridging, which for reasons of space we cannot discuss here, the robustness of indefinite bridging argues in favor of a basic mechanism for bridging that is indifferent to definiteness. This is one feature of our proposed account, to which we now turn.

3.2 Concept-level coherence: The case of bridging

We first show how "basic" bridging phenomena are treated in our model, and then turn to more complex cases. In the interests of readability, we provide a relatively informal presentation. However, all of the ideas can be specified precisely in the cognitive architecture outlined in Section 2, or in any other suitable computational cognitive model of concepts.

We begin with example (11):

(11) My car isn't drivable. **A/the fuse** is blown.

In (11), *fuse* (definite or indefinite) is readily understood as a fuse in the speaker's car. Here is how we model this bridged interpretation in our framework.

Mention of the speaker's car induces the construction of an instantiation **car** of the concept CAR, including salient or relevant features of cars and their various inter-relationships (in the hearer's concept). A simplified representation of the simplest possible instantiation created from the first sentence in (26) is shown in Figure 1.

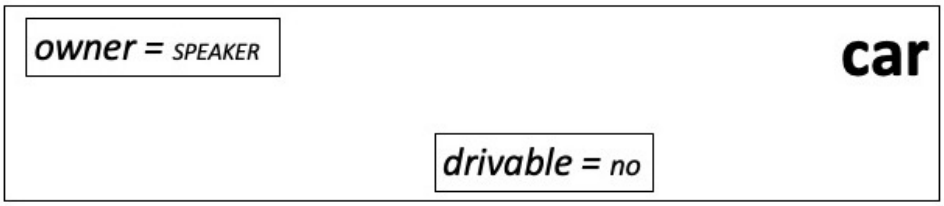

Figure 1: Simplest possible instantiation

Typically, however, the speaker will instantiate other relevant or salient features, depending on a range of factors (De Groot 1983; Neely 1977, 2012). Different speakers may instantiate different features on first hearing some word, and a given speaker may instantiate different features on different occasions. One possible richer instantiation is shown in Figure 2, where we include cartoon probability distributions to indicate variables/features whose values are not precisely known. For example, "gas level" would have a distribution over values corresponding to the state of the tank (full, half-full, etc.). The feature value that is known from the initial sentence in (11) —that the car is not drivable—is encoded in the instantiation as a variable value, rather than a distribution.

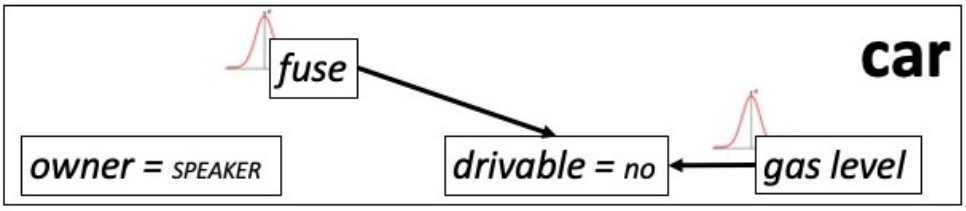

Figure 2: Example initial instantiation

After the hearer responds to the initial sentence of (11) by creating the instantiation in Figure 2, she processes the second sentence. The occurrence of the new NP *a/the fuse* is an instruction to the hearer that an instantiation of FUSE is needed for continued interpretation. In our toy example,

an instantiation of FUSE is already present. The hearer simply utilizes this, and updates the relevant variable value as shown in Figure 3.[7]

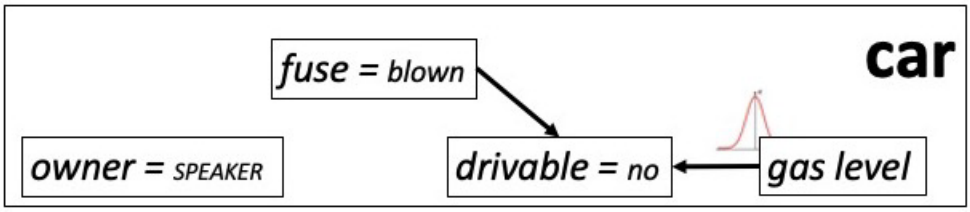

Figure 3: Example final instantiation

The hearer now has a single integrated representation of the content of the two-sentence sequence, where the blown fuse belongs to the **car** instantiation produced in response to the initial sentence. That is, we have a simple case of bridging, as well as a highly concept-level coherent sequence.

Figure 2 represents the assumption that FUSE is relevant or salient enough to be included in the initial instantiation of CAR. While possible, this is not very plausible except in specific conversational settings. If the interpreter does *not* instantiate FUSE when **car** is generated, then she might—depending on exactly what is currently salient—construct something like the instantiation in Figure 4:

Figure 4: Alternative initial instantiation

In this case, upon encountering *a/the fuse* in the second sentence, the hearer is prompted to newly instantiate the concept FUSE in working memory. To derive the bridged interpretation, this instantiation must, again, be represented as a feature of the current **car** instantiation, as illustrated in Figure 5. Explaining why this occurs is the core piece of our account of bridging; we now turn to this explanation.

[7] That value update also triggers inferences about other variables in the instantiation, based on the relevance relations encoded in the instantiation of **car**. In this particular case, the only node that might be updated (*drivable*) already has a known value, so no further inference occurs.

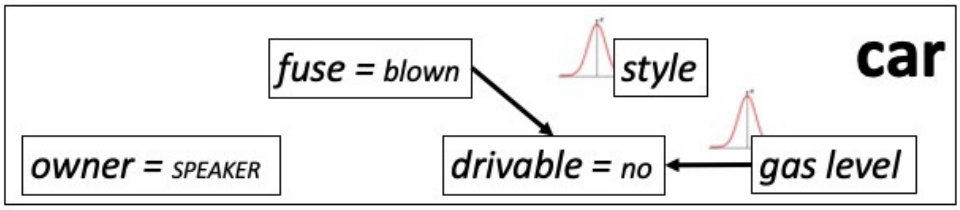

Figure 5: Alternative final instantiation

Our account relies on spreading activation, a standard cognitive model for many types of search, inference, or prediction (Collins and Loftus 1975). Roughly, the idea is that mental content can be "activated," and that activation can spread out along paths determined by the individual's knowledge, thereby activating other pieces of knowledge. Perhaps the best-known example of spreading activation is semantic priming: if participants are shown money, then they (are more likely to) disambiguate the sentence "I'm going to the *bank*" to refer to a financial institution; if they are shown water, then they (are more likely to) disambiguate it to the side of a river (see e.g. Meyer & Schvaneveldt 1971, Neely 1977, 2012). The spreading activation explanation of semantic priming is that the initial images activate the corresponding concepts in participants' minds, and that activation "spreads" to semantically related concepts, but not to unrelated ones. When trying to interpret the sentence, people are more likely to utilize the concept that was already activated.

Importantly, the pathways for activation are provided by existing conceptual structure, both relations between concepts and concept-internal relations. Moreover, it is not only the "end point" concepts that are activated, but the relations themselves: informally put, showing a person a picture of money and uttering the sentence *I went to the bank* will also increase activation of the semantic connections *between* money and banks.[8]

In the probabilistic graphical model approach employed here, graphical edges correspond to the paths along which activation spreads, and which are themselves activated in this process. We contend that spreading activation can explain how connections between mentioned entities are built. These connections are simply the manifestation of spreading activation along connections of informational relevance in the hearer's semantic memory (shaped by some additional factors, such as prior salience/activation of a feature). Spreading activation thus yields concept-level coherence:

[8] Spreading activation is an intriguing metaphor that can be used to explain a number of different psychological findings. That explanatory breadth raises worries that it is only a metaphor, not an actual theory (see Dacey 2016). The PGM approach, however, provides computational grounding for this idea. Evidence for neural implementations of graphical model representations (e.g., Lee and Mumford 2003; Tervo, Tenenbaum, and Gershman 2016) suggests a mechanism by which spreading activation could instantiate features and relationships that are not explicitly stated.

it results in representations that attach new information to an existing representation, yielding a coherent model. Moreover, a representation generated in this way connects the entities mentioned via the strongest connections available, and so the most sensible integration—in this case, the most sensible bridge—automatically occurs. When we have mention of both *car* and *fuse*, both concepts are activated, and the relation between them will have a higher level of activation than, say, the relation between fuses and houses. There is no reason, then, for the interpreter to think about fuses in the speaker's house, and to try to relate one of *those* fuses to the speaker's car. No special principles of coherence are required to explain this; the explanation is provided directly by the cognitive mechanisms that yield an integrated (i.e., concept-level coherent) representation.

The spreading activation model also accounts for how a speaker "chooses" among competing potential antecedents for bridging. Consider (12), where the first sentence contains two possible antecedents:

(12) My car is in the repair shop. A/The fuse is blown.

When the interpreter hears the first of these sentences, the concepts CAR and REPAIR SHOP are both activated in the interpreter's semantic memory, and instantiated in her working memory. Let's assume that, although both cars and repair shops may have fuses, this information is not included in either instantiation. What then makes *car* the more likely antecedent? Our assumption is that for most people, FUSE and CAR are more closely related than FUSE and REPAIR SHOP. Consequently, when the three concepts, FUSE, CAR and REPAIR SHOP are all activated, activation of a relation between the first two will arise more quickly or strongly than activation of a connection between FUSE and REPAIR SHOP. And once a way of integrating the new information to the current instantiation is available, processing of *fuse* can stop.

Further possibilities of the framework are illustrated by (13).

(13) I have to move out of my apartment. The doorman is harassing me.

Here, *doorman* is interpreted as "doorman of the building containing my apartment," and so bridging requires the interpolation of an entity that has not been explicitly mentioned. On our model, after processing the first sentence, the interpreter will have activated the concept APARTMENT and created an instantiation of it. Activation of APARTMENT plausibly already leads to increased activation of APARTMENT BUILDING, although not necessarily reaching the level

necessary for instantiation. Now the interpreter hears *the doorman.* This NP activates the DOORMAN concept, and activation from this concept spreads, plausibly further increasing activation of APARTMENT BUILDING sufficiently that this concept, and the structure of relevance relations that connect it to both DOORMAN and APARTMENT, become instantiated. The interpretation of *doorman* as "doorman of the building that contains my apartment" naturally occurs in the process of incorporating the content of the second sentence into the existing instantiation.

4. Concept-level coherence

The previous section illustrated at a small scale how concept-level coherent interpretations arise, and how coherence at this level generates bridged interpretations of NPs. In this section, we take a step back and describe the notion of concept-level coherence at a higher level.

As is evident, concept-level coherence is a strictly *cognitive* notion of coherence. Moreover, unlike the view of coherence offered by SDRT and related frameworks, it is a *process*-based notion rather than an output-based notion. SDRT considers a discourse to be coherent just in case it is possible to attach each discourse segment, at the point at which it occurs, to the discourse structure, by some coherence relation. We can thus in principle determine the coherence of a discourse by observing the final form of the structure assigned to it: if all segments are attached, and all relations are allowable, the discourse is coherent. In contrast, concept-level coherence cannot be "read-off" the final form of the (complex) instantiation; what will have made the discourse coherent, in this sense, is whether instantiation construction was able to proceed utilizing highly activated interconnections between mentioned entity-types (and event-types, a point we return to in a moment). The more the process requires the creation of novel connections, the less concept-level coherent it will be.

As is clear from the latter point, concept-level coherence also differs from discourse-based notions of coherence in that it is a graded phenomenon, rather than a dichotomous one. This predicts that (to the extent that judgments of coherence reflect concept-level coherence), these judgments should be graded rather than binary. In our account, the degree to which a discourse is coherent will depend on the nature and number of interconceptual relations which are utilized in constructing the instantiation.

Standard notions of coherence have been used not only to capture speaker/hearer intuitions of "good" and "bad" sequences, but also to explain how coherence-building results in enrichment of the content conveyed by a sequence of utterances. Our discussion of bridging illustrates in detail how this arises for NP interpretation. An important difference between our model of how this occurs and extant models is that we take enrichment to be a *consequence* of instantiation construction, rather than, as in other approaches, a *prerequisite* to establishing a coherence relation. Consider, for example, (14), from which the hearer learns indirectly that the room in question has a chandelier:

(14) She looked into the room. The chandelier sparkled brightly.

Clark 1975 posits that this information is "learned" because it must be *supposed* in order to create the relevant bridge. In our account, because the output of interpretation is an integrated world model, the result just *is* an instantiation of a room containing a chandelier. The information is inferred from the representation, rather than being presupposed in order to generate the representation.

Another positive feature of this process-based account of coherence is that it easily accounts for the non-monotonicity of coherence-driven inferences (including bridging). Non-monotonicity is illustrated by the contrast between (15) and (16):

(15) a. I went to a concert last weekend. It was great.
b. But the back-up singers had a really weird choreography.
c. It was quite disturbing.

(16) a. I went to a concert last weekend. It was great.
b. The orchestra was superb, especially the strings section.
c. They played some wonderful Bach.
d. But the back-up singers had a really weird choreography.
e. It was quite disturbing.

Our model easily explains the fact that introducing more information ((16)b,c) undermines the coherence of the previously coherent sequence (15). After the initial sentence in each sequence, the interpreter instantiates her concept of CONCERT without much further specification, as nothing in the context has provided any information about the concert in question. On hearing *back-up singers* in (15)b, she instantiates that concept, and its activation spreads, plausibly activating a sub-type of CONCERT – the kind with back-up singers – along with other related concepts (perhaps

GUITAR, ROCK BAND, etc.). All of the sentences can be easily integrated into a single instantiation. In the case of (16), however, the discourse provides further information about the concert: there was an orchestra, and they played Bach. This information is incorporated into the instantiation, and the resulting inferences lead the hearer to an instantiation in which it is highly probable that this is a classical orchestral concert. When the hearer encounters *the back-up singers* in (16)d., that concept becomes activated and instantiated, but now there is no way to integrate the information into the existing instantiation (since a rock concert is (almost) never a classical orchestra concert). Either the hearer must integrate the concept instantiations anyway (resulting in a *very* unusual concert) or maintain two concert instantiations (resulting in incoherence over the sequence of utterances). In either case, sequence (16) exhibits low concept-level coherence.

In our discussion, we have focused on the (relatively) easy case of concepts for basic entities, like cars and windshields. In principle, though, the same understanding of concept-level coherence can be extended to much more complex conceptual representations, including concepts for types of events and states. Quite plausibly, proposition-level coherence relations, including many of the relations posited in SDRT and in Kehler's work, describe types of relations that can hold at the conceptual level. However, in our model we allow relations at *all* levels of conceptual structure to contribute to coherence building, and hence to interact. Entity-level relations may feed the construction of higher level event-level relations, as well as vice versa. To illustrate this interaction, consider example (17):

(17) Jane was at the playground. She played on the model firetruck.

Plausibly, for many hearers, *model firetruck* is not a common value for the "equipment" variable of their PLAYGROUND concept. But undoubtedly, playing events are represented in that concept. So as the listener builds a representation of a playing event, it will naturally be located in the playground (or in the state of being in a playground) already represented. And as the model firetruck is now a location for the playing, then it too must be located in the playground, and the content of the second sentence serves as an elaboration of the content of the first. Through this interaction of event-concepts and entity-concepts, the hearer arrives at a representation in which the model firetruck is in the playground. (And since this information may be used to update the concept stored in semantic memory, the hearer has now learned about this new kind of playground equipment.)

We conclude with a final observation about knowledge representation and discourse coherence. As we have noted more than once, all coherence theorists acknowledge the important role that world knowledge must play in any theory of coherence. The coherence theories that have dominated the literature to date assume (either explicitly or implicitly) that knowledge is represented propositionally. It is therefore natural to assume that coherence relations will arise through propositional reasoning, and will reflect knowledge of relations between propositions. However, if one begins, as we do, from the view that much background knowledge – as well as the output of interpretation – is represented by conceptual structures, it is natural to model coherence as resulting from integration of these structures. Propositional information can of course be retrieved or generated from these structures; but the processes that operate on these structures, including processes of learning and inference, do not necessarily involve the manipulation of propositions. We believe that conceptual structures are a cognitively more plausible foundation for knowledge representation; but the more important point with which we conclude is the deep connection between models of knowledge representation and notions of coherence.

Appendix: Formal model of instantiation construction with spreading activation

On the concept side, assume a hearer has semantic knowledge represented as a graphical model G (not necessarily connected), where some nodes of G are variables/concepts that themselves have graphical model structure. We further assume a relevance function for each variable $R_X(Y)$ that provides the relevance of variable Y to variable X (i.e., how readily one thinks of Y, given that one is thinking of X). The relevance functions need not be symmetric; that is, possibly $R_X(Y) \neq R_Y(X)$. In general, the relevance functions can vary across conversational contexts and hearer goals, but we assume that they are stable within a particular discourse. We require that the qualitative structure of relevance functions (i.e., the variables for which they are non-zero) be identical with the informational structure represented in the graph component of G, but the quantitative features need not be the same (e.g., X and Y might be highly correlated but not strongly relevant for one another.) There are standard experimental techniques to estimate G (Goodman et al. 2015; Danks 2014; Rehder 2003b) and the R_X functions (Balota and Lorch 1986; De Groot 1983).

On the language side, we assume a mapping from uttered nouns to concepts (i.e., variables/nodes) in G. We also assume that there is a mapping from various linguistic templates to operations on representations. For example, the words "*A* is *b*" or "*A* has *b*" prompt the operation of assigning the value *b* to *A* (i.e., predication). In general, these mappings are often triggered by verbs. Viewed from a slightly different perspective, these operations can be understood as realizations of the surface meaning of an utterance. We do not attempt to specify a general mapping from particular templates (or verbs) to cognitive operations, as that would involve giving a full theory of meaning, rather than our present narrower focus on bridging phenomena.

Let $Act(X, t)$ be the activation level of node X in G at time t;[9] if $Act(X, t)$ exceeds a threshold τ, then the concept X is instantiated as node $\mathbf{x}$ in $\mathbf{W}$, the working memory of the hearer. The instantiation(s) in $\mathbf{W}$ thus form a (not necessarily connected) graphical model $\mathbf{G}$. τ will typically depend on contextual factors including the hearer's attentional focus, but we assume that it is constant for the duration of a few utterances. We are deliberately agnostic about whether working memory is neurally distinct from episodic or other short-term memory structures. We assume only that $\mathbf{x}$, the working memory instantiation of X, is independently manipulable from the semantic

[9] For convenience, we assume discrete time, though everything in this Appendix could, at cost of increased complexity and underdetermination, be rewritten in terms of continuous-time functions.

knowledge representation of X. However, we assume that there is a persistent link between **x** and X; colloquially, **x** is marked as an instance of X.

When the hearer processes a noun 'x', the activation of X (as given by the language function) is increased significantly above τ. This step—hearing a word triggers activation of a concept sufficient to bring it to awareness—is standard in essentially all theories of language processing and comprehension (see, e.g., Devereux, et al. 2013 or Mirman and Magnuson 2009 for possible neural accounts of this mapping). There are multiple plausible explanations for exactly how this process might work at lower- or higher-levels of description. If there is an **x** in **W**, then that instantiation receives the hearer's focus. If there is no such **x**, then one is instantiated since $Act(X, t) > \tau$. If the hearer processes additional words that match a template for a cognitive operation (e.g., predication), then that operation is applied to the instantiations in **W**. These elements suffice for generation of the literal meaning of an utterance, but no further inferences. Two types of inferences now occur in parallel: spreading activation to connect together elements of **G** based on relevance relations R_X, and message-passing to propagate information across **G** based on informational probabilities P.

There is substantial evidence for spreading activation within semantic memory: concept activation (even if below τ) causes related concepts to be partially activated via neural firing patterns along connections between relevant content (Fuster 1997; Tulving 1983; Baddeley 2012). For example, suppose that X and Y are connected in semantic knowledge. Mathematically, if $Act(X, t)$ increases by δ, then $Act(Y, t+1) = Act(Y, t) + \delta R_X(Y)$.[10] If $Act(Y, t+1) > \tau$, then **y** will be instantiated in **W** (or if a **y** already exists, then it receives attentional focus), and the $X - Y$ information connection in G is also instantiated in **G** (in **W**). Note that this process implies that "sufficiently relevant" features F of a concept X—that is, those with sufficiently high $R_X(F)$—will automatically be instantiated when X is instantiated. Similarly, features that were already somewhat salient in the discourse (and so had higher prior activation) are more likely to be instantiated alongside X. It is possible to have instantiated concepts with no features (e.g., "blank" concepts often used in psychological experiments), but this rarely occurs in everyday life.

Activation iteratively spreads: if $X \rightarrow Y \rightarrow Z$ in G, then $Act(Z, t+2) = Act(Z, t+1) + \delta R_X(Y) R_Y(Z)$, potentially moving $Act(Z, t+2)$ above τ. This spreading activation process thus determines the

[10] This functional form implies non-trivial constraints on the scales used for $Act()$ and $R()$ functions.

elements—nodes and edges—instantiated in **G**. Over time, content activations will tend towards zero; that is, in the absence of other input, $Act(X, t+1) < Act(X, t)$. However, if the hearer maintains consistent focus on content X, then activation will continue to spread out through the network. As a result, a (potentially quite distal) node S can eventually have $Act(S, t+k) > \tau$ for relatively large values of k.

In parallel, cognitive operations can change the (probability distribution over) values of nodes in **G**. That is, the elements of **G** are not simply connected together via activation, but information about the values of those variables then propagates across those connections. Whenever distributions change, standard graphical model message-passing algorithms convey information throughout elements of **G** (see Lee and Mumford 2003; Tervo et al. 2016; and references therein for neural evidence of message-passing as a distinct operation from spreading activation). For example, if $\mathbf{x} \rightarrow \mathbf{y}$ and **x** is changed to the value a, then **y** will be updated to $P(y \mid \mathbf{x} = a)$.[11] These value-updates are based on the informational relations in **G**, not the relevance relations.

[11] Assuming **y** previously had distribution $P(\mathbf{y})$, which might itself have been the result of previous updating via message-passing.

References

Asher, N. 1993. *Reference to Abstract Objects in Discourse.* Berlin: Kluwer Academic Publishers.

Asher, N. and Lascarides, A. 1998. Bridging. *Journal of Semantics* 15, 83-113.

Asher, N. and Lascarides, A. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.

Clark, H. 1975. Bridging. In R. C. Schank and B.L. Nash-Webber (eds.) *Theoretical Issues in Natural Language Processing.* New York: Association for Computing Machinery, pp. 175–196. Reprinted in P.N. Johnson-Laird, P.C. Wason (eds.) *Thinking: Readings in Cognitive Science*. Cambridge University Press. Cambridge, 411-420, 1977.

Collins, A.M. and Loftus, E.F., 1975. A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407-428.

Dacey, M. (2016). Rethinking associations in psychology. *Synthese* 193(12), 3763–3786.

Danks, D. (2007). Theory unification and graphical models in human categorization. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 173–189). Oxford: Oxford University Press.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Cambridge, MA: The MIT Press.

De Groot, A.M., 1983. The range of automatic spreading activation in word priming. *Journal of verbal learning and verbal behavior*, *22*(4), pp.417-436.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000-2005. *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.

Goodman, N., Tenenbaum, J., & Gerstenberg, T. 2015. Concepts in a probabilistic language of thought. In Margolis & Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–654). Cambridge, MA: MIT Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111* (1), 3-32.

Griffiths, T.L. and Tenenbaum, J.B., 2005. Structure and strength in causal induction. *Cognitive psychology*, *51*(4), 334-384.

Gundel, J.K., N. Hedberg & R. Zacharski 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69:2, 274-307

Hobbs, J. 1979. Coherence and Coreference. *Cognitive Science* 3, 67-90.

Hobbs, J. 1990. *Literature and Cognition*. Stanford: CSLI Publications.

Kehler, A. 2002. *Coherence, Reference, and the Theory of Grammar*. Stanford. CSLI Publications.

Kehler, A. 2015. Reference in Discourse. In Shalom Lappin & Chris Fox (Eds), *The Handbook of Contemporary Semantic theory, 2nd edition*. New Jersy: John Wiley and Sons.

Kehler, A. 2012. Cohesion and Coherence. In C. Maienborn et al. (eds), *Semantics: An International Handbook of Natural Language Meaning*, Berlin: De Gruyter, pp. 1963--1987.

Lee, T.S. and Mumford, D., 2003. Hierarchical Bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434-1448.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome?. *Cognitive Psychology, 32*, 49-96.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. Journal of experimental psychology, 90(2), 227.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. Journal of experimental psychology: general, 106(3), 226.

Neely, J. H. (2012). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In Basic processes in reading (pp. 272-344). Routledge.

Prince, E. 1992. The ZPG Letter: Subjects, Definiteness, and Information-status. In William C. Mann and Sandra A. Thompson (Eds.), *Discourse Description: Diverse linguistic analyses of a fund-raising text*. Philadelphia: John Benjamins, 295-325.

Roberts, C. 2003. Uniqueness in Definite Noun Phrases. *Linguistics and Philosophy* 26:3, 287-350.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology, 35*, 1-33.

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Current Opinion in Neurobiology, 37,* 99-105.