Proceedings of SALT 35: 390–409, 2025

Counterfactual interpretation as search for coherence in a learned model of the world*

Adrian Brasoveanu
UC Santa Cruz

Abstract We propose a semantics for counterfactuals that explicitly connects it to cognitive models of temporal / causal inferences in narrative discourse comprehension. Both narrative discourses and counterfactuals are understood by (i) constructing temporally-indexed sequences of situation models – analog, vector-space representations of meaning grounded in our experience perceiving and understanding the world (our learned model of the world) – and then (ii) enriching these representations with temporal inferences through a temporal coherence-seeking process. The account builds on the model of narrative discourse comprehension in Frank et al. 2003, showing how we can use it to move beyond theoretically primitive notions of counterfactual similarity by computing / constructing a historically-structured similarity relation through this coherence-seeking interpretation process. We formalize a semantics that integrates linguistic input with rich temporal world knowledge, interprets counterfactuals in a graded and context-sensitive manner, and inherently links interpretation to processing time.

Keywords: counterfactuals, situation models, discourse comprehension, vector space semantics, incremental processing of semantic representations

1 Introduction and the informal account

This paper proposes an interpretation for counterfactuals that leverages the same cognitive processes and knowledge bases underlying our highly practiced skill of narrative discourse comprehension. The core idea is that counterfactuals, just like narratives, are understood by continuously trying to fit what is heard or read into our learned model of the world, using our background world knowledge to

* I want to thank Stefan Frank, Jessica Rett, Harm Brouwer, Noortje Venhuizen, Dan Lassiter, Dylan Bumford, Jakub Dotlačil, Boram Kim, Donka Farkas, Chris Barker, the audiences of the UCLA Syntax & Semantics Seminar (May 7, 2025), UCSC S-Circle (May 14, 2025) and SALT 35 (May 20-22, 2025), as well as the SALT 35 anonymous reviewers, for all their comments and questions about this work. My conversations with Jessica Rett, Stefan Frank, Harm Brouwer and Noortje Venhuizen in particular have been extremely helpful and enlightening. I am also grateful to Stefan Frank for sharing his original Matlab implementation of Frank, Koppen, Noordman & Vonk 2003, which was the starting point for the Python implementation of the present account.

make inferences and fill in the gaps. Consider, for example, this elementary-school narrative, first discussed in Stein & Glenn 1979:

(1) "Once there was a little boy who lived in a hot country. One day his mother told him to take some cake to his grandmother. She warned him to hold it carefully so it wouldn't break into crumbs. The little boy put the cake in a leaf under his arm and carried it to his grandmother's. When he got there, the cake had crumbled into tiny pieces. His grandmother told him he was a silly boy."

As we read (1), we build a 'mental movie' consisting of a temporally-ordered sequence of *situation models* (Zwaan 2001; Zwaan & Radvansky 1998; Cohn 2019): a sequence of rich, analog representations of the narrated events unfolding over time. We can think of them as evolving mental images like the drawings in Figure 1.

Figure 1 Depictions suggestive of situation models constructed during the interpretation of discourse (1). (generated with GPT-4o, May 2025)

A key aspect of situation-model construction is our reliance on background world knowledge to go beyond what's explicitly stated. In our story in (1), why is the boy "silly"? The text doesn't explicitly say. We infer it (Noordman & Vonk 2015 a.o.) because our world knowledge includes facts like "cakes are fragile and can crumble if squeezed / not handled carefully." The narrative implies a sequence of events: (i) at t_1, the boy puts the cake in a leaf under his arm (squeezing it); (ii) at t_2, the cake crumbles into pieces. The inference that the boy's action at t_1 caused the crumbled cake at t_2 is not present in the text – it's an enrichment coming from our background understanding / knowledge of the world's temporal dynamics.

This brings us to the *two main proposals* in this paper. *The first proposal* is that discourse interpretation is a dynamic process of aligning the initial 'static', i.e., non-temporal, interpretation of the discourse with our background knowledge of the world's temporal dynamics. This alignment process can be characterized as a 'search' for coherence, but formally, it is an inferential process that uses our background world knowledge – our model of the world's temporal dynamics learned

from experience – to enrich the explicitly stated story with additional inferences. The resulting interpretation is a sequence of situation models that is more causally and temporally 'connected' / coherent. We build extensively on the cognitive model of discourse comprehension in Frank et al. 2003, which itself builds on Golden & Rumelhart 1993; Golden, Rumelhart, Strickland & Ting 1994.[1] The novel theoretical contribution of the present paper is a systematic reconceptualization and restructuring of the main components of the Frank et al. cognitive model as a semantic framework recognizable as such by formal semanticists, where we explicitly define the semantic interpretation in terms of models learned from experience / data, and that can generalize beyond that input experience / data.

Thus, interpretation is a coherence-seeking process in a specific sense: it is not an active search for an end goal state, but a constructive inferential process that increases alignment with our background knowledge of the world's temporal dynamics. And we are not seeking *textual* coherence, but coherence, or increased alignment, of what is explicitly said – what we called the initial 'static' interpretation of the discourse – with our learned model of the world's temporal dynamics.

This inferential enrichment is interpretive work that our cognitive system performs, and as such, it takes processing effort and time. We will therefore formalize semantic interpretation in a way that directly connects processing time with how much the initial 'static' interpretation of the discourse is modified to become more coherent / aligned with our background knowledge of the world's temporal dynamics.

While *compositional* interpretation at the sub-sentential level is undoubtedly part of the cognitive work of discourse comprehension, we are focusing here on a different, equally crucial aspect: the pervasive 'pragmatic intrusion' of world knowledge throughout the process of constructing the overall interpretation of a discourse. In fact, calling this cognitive work 'intrusion' is misleading, as this inferential process is not merely an add-on. For the purpose of using language to better understand the world and more effectively do things in the world jointly with other humans, it is the *main* work of discourse interpretation.

The 'pragmatic intrusion' label is misleading for another reason: it takes for granted a separation between semantics and pragmatics that might simply be a methodological artifact of the mathematical-logic approach we take in formal semantics. This Montagovian approach has the virtue of placing the 'aboutness' of language at the heart of semantic theory: meaning connects linguistic forms with the (non-linguistic) world. But it is not obvious that the relatively clean distinction between semantics and pragmatics that we can make for formal languages also applies to natural language interpretation. In this paper, inferential enrichment is formalized as an essential component of the *semantics* of natural language discourses: it is a

1 See also Venhuizen, Hendriks, Crocker & Brouwer 2022 and references therein for more recent related developments.

core part of the meaning construction process that our cognitive system performs to understand natural language. Interpretation / meaning construction is fundamentally a process that integrates linguistic input with our rich pre-existing world knowledge.

The second main proposal, and the more empirically specific contribution of this paper, is that this coherence-seeking semantics framework extends naturally to the interpretation of counterfactual statements. For example, we could continue the story in (1) with the following counterfactual statement:

(2) [His grandmother told him he was a silly boy. She said that ...] If the boy had carried the cake on top of his head, the cake wouldn't have crumbled.

We propose that interpreting counterfactuals leverages the same coherence / alignment seeking inferential process that underlies the interpretation of narratives. We thus ground counterfactual interpretation in the independently motivated cognitive model of discourse comprehension in Frank et al. 2003. Specifically, we propose that to interpret a counterfactual, the cognitive system constructs a situation model for the counterfactual antecedent that is as similar as possible to the actual context by maximizing (up to a threshold) the alignment / temporal coherence between the counterfactual antecedent and the actual context.

This constructive approach to similarity contrasts with the standard Stalnaker; Lewis semantics for counterfactuals $A > C$, which relies on a vague, theoretically primitive notion of similarity $\leq$ between worlds that resists detailed formalization: $A > C$ is true at $w_@$ iff C is true in the A-worlds most $\leq$-similar to $w_@$.

Historically-structured similarity relations have been proposed to address some of the shortcomings of pure similarity accounts (Thomason 1970; Ippolito 2013; Khoo 2017), but these approaches still leave the similarity relation as an unanalyzable primitive in fundamental respects. In contrast, we leverage the coherence-seeking inferential process underlying narrative discourse comprehension to *construct* the similarity relation. Having a model of the world's temporal dynamics learned from experience will be an essential ingredient of this constructive approach to similarity.

To formalize this approach to the semantics of counterfactuals, what kind of representations and processes do we need? Consider the counterfactual (CF) in (4), which will serve as our running example for the remainder of the paper:

(3) [Context] It's raining at t_1. Bob and Jane are playing computer games at t_2.

(4) [CF] If it had been sunny at t_1, they would have played outside instead at t_2.

To interpret (4) in the context of (3), we require three key components. First, a notion of *similarity / distance between situation models*: we need a way to construct a situation model for the antecedent – where it's sunny at t_1 instead of rainy – that is otherwise as similar as possible to the actual context. Second, we need to

be able to *predict how events typically unfold*: based on the counterfactual sunny situation at t_1, which is otherwise as similar as possible to the actual world, we need to predict what would happen next at time t_2. We will construct this predicted counterfactual situation model at t_2 based on our learned model of the worlds's temporal dynamics. Third, we need a way to draw *plausible inferences*:[2] how likely is the counterfactual consequent given the predicted counterfactual situation model at t_2? For (4) specifically: is it more likely that Bob and Jane play outside (the CF consequent) than that they play computer games (the actual activity)?

Formal semantics, while providing sophisticated tools for compositional interpretation, provides little formal machinery for these kinds of graded, world-knowledge based similarity judgments, predictions and inferences. The main tools are custom definitions of modal accessibility relations, supplemented as needed (often on an example by example basis) by informal pragmatic reasoning.

In this paper, we make formal progress on the three key components outlined above. This progress is guided by two further desiderata. First, we want a framework that can make *up-front similarity predictions*: we want to be formal, systematic, and up-front about the similarity of situations / situation models, how events typically unfold, and what plausibly follows from what. This should be possible to state before encountering any particular example discourse or counterfactual, moving beyond informal talk about vague modal accessibility relations and their pragmatics and towards theories that can make precise, yet still graded (numerical) predictions.

Second, we want a *semantic theory that can inherently make processing predictions*, i.e., one that can predict measurements of human verbal behavior. This contrasts with approaches that first build a competence-level semantic theory and then face the enormous burden of designing a separate, formally explicit linking theory to connect it to experimental data. In principle, this separation of concerns seems like a good idea. In practice, the burden of developing a formal linking theory is so great that it never truly happens. Our goal is to build the foundations of a formal linking theory as an integral part of the semantics, effectively rebuilding semantic theory from scratch with an eye towards processing predictions.

2 Sentence meaning in a model of the world learned from experience

The foundation of our approach lies in how we represent the meaning of individual sentences and, crucially, how these meanings are grounded in a learned model of the world. We conceptualize understanding a sentence as the process of constructing a situation model – a mental representation of the described situation. This is akin to mentally 'simulating' the state of affairs the sentence depicts, thereby understanding

2 See Kaufmann 2015; Lassiter 2017 a.o. for probabilistic approaches to conditionals and modality.

the conditions under which the sentence would be true. Thus, in our framework, situation models are non-linguistic, truth-conditional, and critically, *analog* mental representations. They are: (i) similar to the result of actually perceiving the described situation; (ii) disquotational, in that they 'simulate' the conditions under which the sentence (construed as a description of a situation model) is true; and (iii) importantly, this 'simulation' possesses analog properties that enable us to make certain kinds of inferences by direct inspection, rather than relying solely on symbolic manipulation (Zwaan 2008; Frank & Vigliocco 2011; Noordman & Vonk 2015 a.o.).

2.1 The experience / training data for our learned model of the world

To formally develop and test our model, we utilize an example microworld, adapted from Frank et al. 2003, centered around the activities and states of two characters, Bob and Jane. In this microworld, a situation at any given time t is exhaustively specified by the truth or falsity of fourteen atomic propositions, listed in Table 1.

1.	SUN	The sun shines.	8.	J_COMP	Jane plays a computer game.
2.	RAIN	It rains.	9.	B_DOG	Bob plays with the dog.
3.	B_OUT	Bob is outside.	10.	J_DOG	Jane plays with the dog.
4.	J_OUT	Jane is outside.	11.	B_TIRED	Bob is tired.
5.	SOCCER	Bob and Jane play soccer.	12.	J_TIRED	Jane is tired.
6.	HIDEnSEEK	Bob and Jane play hide & seek.	13.	B_WINS	Bob wins.
7.	B_COMP	Bob plays a computer game.	14.	J_WINS	Jane wins.

Table 1 The 14 atomic propositions in our microworld (from Frank et al. 2003).

This simple microworld is rich enough to exhibit complex temporal dynamics and allow for the construction of non-trivial counterfactuals. To learn the regularities of this microworld, we use a dataset of 250 temporally-ordered situations as our experience / training data, shown in Figure 2. Think of this dataset as a sample 'movie' taking place in the microworld and consisting of 250 temporally-ordered 'frames.' Each frame shows a snapshot of the activities and states of Bob and Jane.

In Figure 2, the data is organized as a matrix with 14 rows, one for each atomic prop(osition), and 250 columns, one for each time step. The columns represent *situations*: complete specifications of the world at a given time t with respect to the 14 binary features, i.e., an assignment of truth values $\{1,0\}$ to the 14 atomic props. The rows represent the *truth conditions* of each atomic prop across time: the set of times t where the atomic prop is true (1) or false (0). Thus, the rows provide the data from which we can learn *situation models* / meanings for each atomic prop.

Setting up a microworld and generating this kind of dataset in advance allows us to be up-front about our predictions regarding what constitutes similar situations,

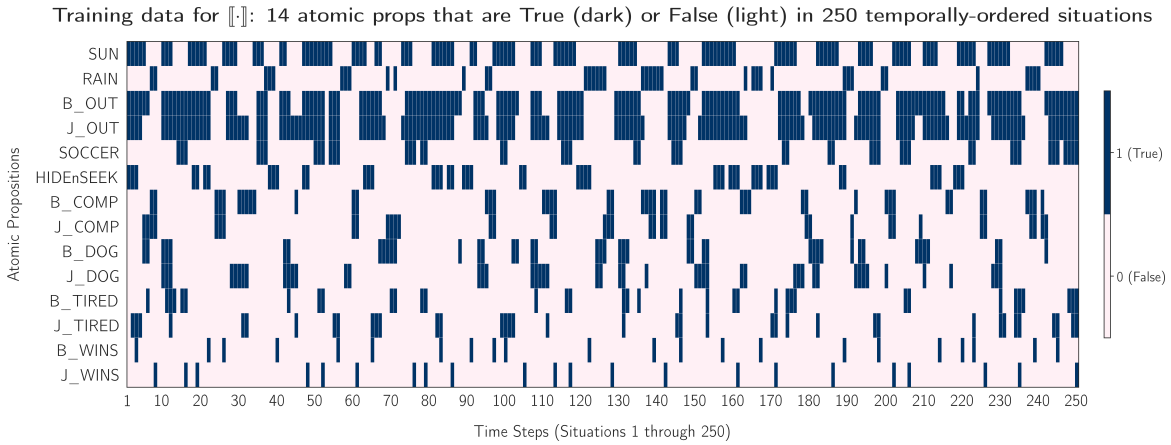

Figure 2 Training data for the interpretation function $[\![\cdot]\!]$: 250 temporally-ordered situations characterized in terms of 14 binary features / atomic props that are either True (dark) or False (light) in any given situation.

how events typically unfold, and what plausibly follows from what, before analyzing any specific discourses or counterfactuals.

2.2 Learning the interpretation function $[\![\cdot]\!]$: two kinds of world knowledge

The dataset in Figure 2 could serve as a model for propositional tense logic (Prior 1955), as each column is an assignment of truth values to the 14 atomic props at a given time t. We will instead focus on the row-wise patterns, and use this 250-frame 'movie' as experience, or empirical grounding, for learning the interpretation function $[\![\cdot]\!]$, i.e., for learning a model of the world. This 14-by-250 binary dataset encodes *two kinds of world knowledge* that we aim to learn.

On one hand, we have *non-temporal / within-time-step constraints*, which govern how likely the states and activities of our two characters Bob and Jane are at a single point in time, and which of these states and activities can or cannot co-occur at any given time t. For example: (i) it can't be both sunny and rainy simultaneously; (ii) being sunny has a higher base probability than being rainy – across our 250 situations, SUN is 1 more often than RAIN is 1; (iii) Bob and Jane can only play soccer if they are outside – in our 250 situations, SOCCER is 1 only if both B_OUT and J_OUT are 1; (iv) they can only play computer games if they are inside, e.g., B_COMP is 1 only if B_OUT is 0; (v) they can only perform one primary activity at a time, e.g., SOCCER and HIDEnSEEK cannot be 1 at the same time; (vi) one of them can win only when playing soccer, hide & seek, or both play computer games; (vii) Bob and Jane are more likely to be in the same place (inside or outside) than in

different places, and to engage in the same activity across our 250 situations.

Thus, the non-temporal constraints govern the *joint distribution* of the 14 binary features / atomic props at any given time t: how likely each of these features is, and how (in)compatible these features are with each other in any given situation / at any give time. The non-temporal constraints include both 'hard' constraints, e.g., SUN and RAIN cannot both be 1, and 'soft' constraints / probabilistic tendencies, e.g., Bob and Jane are more likely to be in the same place. Situation models – our analog representations of the truth conditions of the atomic (and, we will soon see, non-atomic) props – will encode these non-temporal constraints.

On the other hand, we have *temporal / across-time-step constraints* which govern how events typically unfold, i.e., how situations transition over time. These constraints tell us how our 14 binary features, i.e., the states and activities of Bob and Jane, typically evolve from one time step to the next. For example: (i) Bob and Jane typically stop playing a game after one of them wins; (ii) a game can only be won if it was played in the immediately preceding time step(s); (iii) whoever is tired is less likely to win at the next time step; (iv) Bob and Jane are more likely to stay where they are than to change location, unless the weather changes.

Given the training data in Figure 2, our central aim is to *learn the interpretation function* $[\![\cdot]\!]$ – what we described informally as our learned model of the world. The interpretation function $[\![\cdot]\!]$ will do two things for us, tracking the two kinds of world knowledge we discussed above. One one hand, $[\![\cdot]\!]$ will map atomic and (via composition) non-atomic props to 'static' / non-temporal *situation models* that encode their meaning. On the other hand, $[\![\cdot]\!]$ will map discourses to *temporally-ordered sequences* of situation models that not only represent the explicitly-stated 'static' propositions, but enrich them to increase their alignment with the temporal dynamics of the microworld instantiated by the dataset in Figure 2. Thus, the two components of our learned world model / our interpretation function $[\![\cdot]\!]$ are:

(5) *Situation models* $[\![p]\!]$: vector representations of the meanings of each atomic proposition p. They encode an approximation of the *non-temporal* 'hard' (logical incompatibilities) and 'soft' (probabilistic tendencies) constraints observed within individual time steps in the training data.

(6) *Temporal dynamics model (transition matrix W)*: this component encodes an approximation of the temporal 'hard' and 'soft' constraints governing how situations transition from one time step to the next.

While various learning algorithms can be used, we follow Frank et al. 2003 here:

(7) We use *Self-Organizing Maps* (SOMs, a type of competitive neural network; Kohonen 1995) to learn situation models for atomic propositions. SOMs / situation models are continuous, distributed representation of the input binary truth-conditions, with meaningful semantic geometry / analog properties.

(8) We use *Markov Random Fields* (MRFs; Golden et al. 1994) to learn the world temporal dynamics, which will be a transition matrix W.

The specific details of these learning algorithms are less important than the properties of the learned representations that they generate. Therefore, in what follows, we focus on (i) the analog nature of the SOM-learned situation models, which directly supports cognitive operations crucial for interpretation, e.g., drawing plausible inferences, and (ii) the ways the transition matrix W supports computing similarity between meanings, predicting future events, and provides a basis for quantifying the processing time needed for similarity computations.

2.3 Learned non-temporal propositional meanings: analog situation models

The SOM component of the learned interpretation function $[\![\cdot]\!]$ assigns a situation model to each of the 14 atomic propositions in our microworld, visualized in Figure 3.

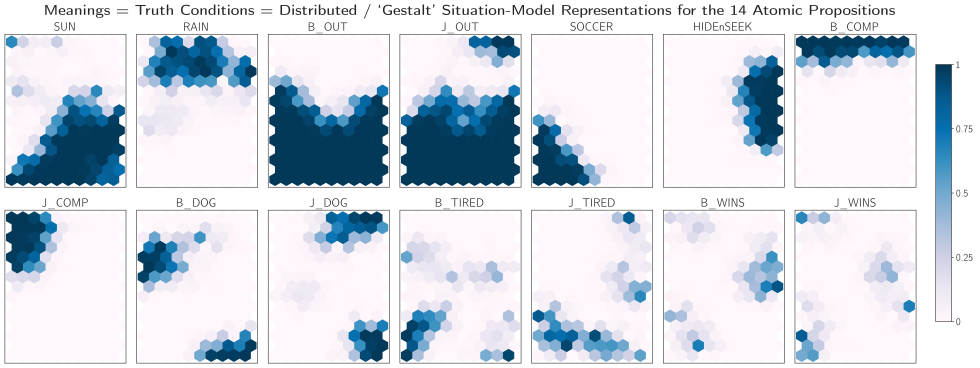

Figure 3 Learned situation models (SOM activation vectors) for the 14 atomic propositions. Each SOM vector represents the truth conditions for a proposition as a distributed / Gestalt pattern of activation across a 150-unit (15×10) hexagonal grid. Darker areas indicate higher activation.

These are approximate mental representations of the meanings / truth conditions of the atomic props that encode both (i) 'hard' non-temporal constraints, i.e., what truth conditions in formal semantics would encode, e.g., the logical incompatibility between SUN and RAIN, and (ii) 'soft' probabilistic non-temporal constraints, e.g., Bob and Jane are more likely to be outside when it's sunny than when it's rainy.

Formally, sentence meanings / lexical meanings / situation models for the 14 atomic propositions are the 14 SOM-learned vectors $[\![\text{SUN}]\!], [\![\text{RAIN}]\!], \ldots$ plotted in Figure 3. These are vectors in a high-dimensional space $\Sigma = [0,1]^n$, where $n = 150$ in our implementation. Each vector is represented as a 15-row $\times$ 10-column 2D map, consisting of 150 hexagonal cells, or units, or neurons, where each unit / cell / neuron has a continuous activation value in the interval $[0,1]$.

The 2D maps are convenient representations of the 150-component meaning vectors that enable us to easily visualize their analog properties – but fundamentally, we think of sentential meanings as 150-dimensional vectors, i.e., points in the $\Sigma = [0,1]^{150}$ space. Each of the 150 units / cells / neurons can be thought of as a *microfeature* of the meaning of the atomic proposition. Microfeatures are not interpretable in isolation. Rather, sentential meaning arises from the overall pattern of activation across all 150 microfeatures. That is, the SOM vectors are *distributed / Gestalt* representations of truth conditions: there is no one-to-one correspondence between a row of 250 binary truth values from the training data in Figure 2 and the 150 microfeatures of the corresponding SOM vector.

While the SOM-vector representations might superficially resemble language-model embeddings, they are fundamentally different: they are *truly semantic* in that they encode *world / non-linguistic patterns* derived from the co-occurrence of binary (atomic prop) features in the situations depicted in Figure 2 (our training 'movie'). Our distributed meaning representations are truly about the world, not just distillations of word co-occurrence patterns from linguistic corpora.

The SOM-learned situation models possess *semantic geometry, i.e., analog properties*: non-temporal 'hard' (truth-conditional) and 'soft' (probabilistic) world knowledge can be directly read off the SOMs. Think of SOMs as maps of mountain ranges: higher regions are plotted as darker SOM cells. The *prior / base probability* of a proposition's occurrence in the microworld, e.g., how often it's sunny, is approximated by the average height of the mountain range plotted on a SOM (the mean of the SOM activation vector). The *likelihood of co-occurrence* of two atomic props / situation features, e.g., the mutual compatibility of SUN and B_OUT, is approximated by the degree of overlap or separation of their mountain peaks (areas of high activation) on their SOMs. A map where we have a *smaller area with peaks* (a smaller dark area) indicates more specific information (fewer compatible situations / microworld states), and hence often corresponds to a lower prior probability. A map with a *larger peaky area* indicates lower informational content or weaker evidence, being compatible with a wider range of situations / microworld states, and therefore having a higher prior probability. Finally, *similarity between meanings*, e.g., $[\![\text{SOCCER}]\!]$ and $[\![\text{SUN}]\!]$, which are similar because soccer is an outdoor activity, which correlates with sunny days, can be naturally computed as a distance in situation-model space Σ, e.g., Euclidean or gradient support as defined in §3 below.

2.4 Learned temporal dynamics of the world: the transition matrix W

The second key component learned from the training data is the *temporal dynamics model*, formalized as an $n \times n$ transition matrix W, where $n = 150 =$ dimensionality of our situation-model vectors. This matrix, visualized as the heatmap in Figure 4,

complements static / non-temporal situation models. It encodes the learned temporal contingencies and associative strengths between situation models across adjacent time steps that are crucial for the interpretation of narratives and counterfactuals.

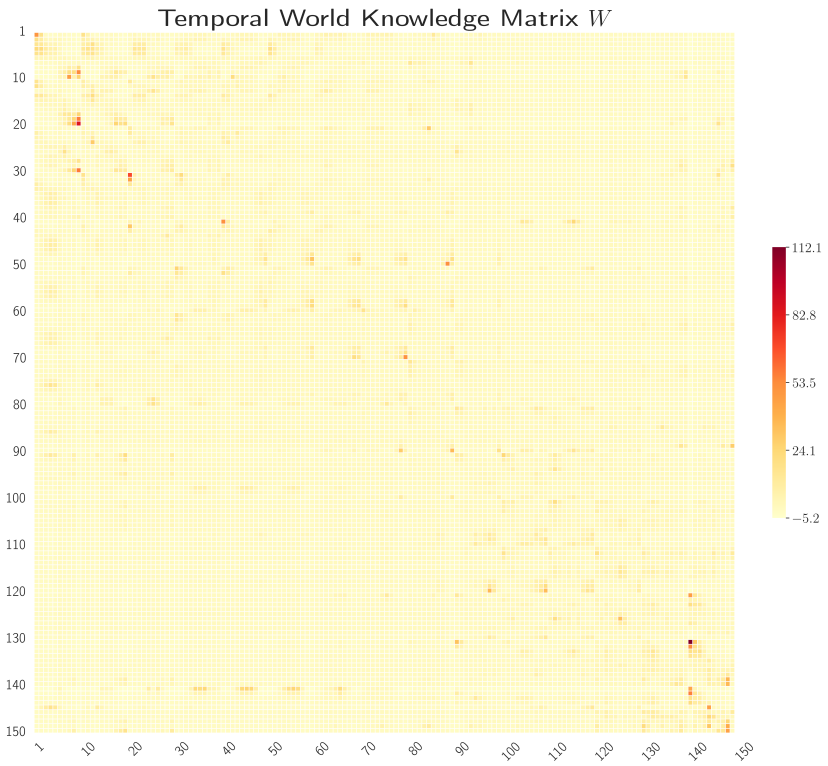

Figure 4 The learned temporal transition matrix W (150×150). Entries w_{ij} represent the learned association strength (covariance) between the activation of microfeature i at time t and microfeature j at time $t+1$. Warmer colors (red) indicate stronger positive associations, cooler colors (light yellow) indicate weak or negative associations.

Each entry w_{ij} in W quantifies how likely (if positive / facilitatory) or unlikely (if negative / inhibitory) it is that microfeature j is active at $t+1$ given that microfeature i was active at t. W is estimated based on temporal co-occurrence patterns (mean lag-1 covariance) in the sequence of 250 time steps in our training data. The transition matrix W captures not only causal connections, but all temporal correlations: what's likely to happen at time t given what's happening at neighboring times $t-1$ and $t+1$. W is fundamental for predicting how events typically unfold, a crucial component for interpreting both narratives and counterfactuals.

3 Sentence-level compositionality: static meanings for non-atomic props

With the lexical meanings (SOM vectors) for atomic props in hand, we can recursively assign meanings to non-atomic props built with propositional operators using vectorial versions of fuzzy logic operations (Zadeh 1975).

The meaning / situation model / probabilistic truth-conditions for the negation of a proposition φ is computed by element-wise subtraction from 1, as shown in (9). Negation effectively inverts the activation pattern of $[\![\varphi]\!]$. The situation model for the conjunction of two propositions φ and ψ is computed using the element-wise Hadamard product, as shown in (10). Conjunction results in a situation model where activation is high only in regions where both $[\![\varphi]\!]$ and $[\![\psi]\!]$ are highly active. Finally, the situation model for the disjunction of φ and ψ is computed by the De Morgan laws translated into vectorial operations, as shown in (11). Disjunction results in high activation in regions where either $[\![\varphi]\!]$ or $[\![\psi]\!]$ (or both) are highly active.

(9) Negation: $[\![\neg\varphi]\!] = 1 - [\![\varphi]\!]$ (element-wise subtraction)

(10) Conjunction: $[\![\varphi \wedge \psi]\!] = [\![\varphi]\!] \odot [\![\psi]\!]$ ($\odot$: element-wise/Hadamard product)

(11) Disjunction: $[\![\varphi \vee \psi]\!] = [\![\neg(\neg\varphi \wedge \neg\psi)]\!] = [\![\varphi]\!] + [\![\psi]\!] - [\![\varphi]\!] \odot [\![\psi]\!]$

The probabilistic nature of SOM-based situation models mentioned in §2.3 above that can be read off their semantic geometry is formally captured as follows:

(12) Prior/base prob. of a prop φ is average activation of situation model $[\![\varphi]\!]$:
$\mathbf{P}(\varphi) = \text{mean}([\![\varphi]\!]) = \frac{\sum_{i=1}^{n} [\![\varphi]\!]_i}{n}$, where $n = 150 =$ number of SOM units

Putting together the definitions of probability and conjunction, we can define a notion of *gradient support* of a proposition ψ by a proposition φ, which is the *conditional probability* of ψ given φ:

(13) Conditional probability / gradient support: $\mathbf{P}(\psi \mid \varphi) = \frac{\mathbf{P}(\varphi \wedge \psi)}{\mathbf{P}(\varphi)} = \frac{\sum_{i=1}^{n} ([\![\varphi]\!] \odot [\![\psi]\!])}{\sum_{i=1}^{n} [\![\varphi]\!]}$

Thus, we are able to compute conditional probabilities directly from the vectorial meaning representations of propositions, showing that our situation-model space Σ inherently supports a notion of *plausible inference*, essential for counterfactuals.

4 Discourse meaning I: trajectories in situation-model space

Having established how atomic and non-atomic propositions φ are mapped to situation models, i.e., SOM-based vectors $[\![\varphi]\!] \in \Sigma = [0,1]^{150}$, we now turn to the meaning of multi-sentence discourses. We define the meaning of a discourse $D = \varphi_1; \varphi_2; \ldots; \varphi_T$ of length T[3] as a *temporally-ordered sequence of situation models*

3 We discuss the dynamic conjunction / sequencing operator ';' below.

$\bar{X} = \langle X_1, X_2, \ldots, X_T \rangle = \langle [\![\varphi_1]\!], [\![\varphi_2]\!], \ldots, [\![\varphi_T]\!] \rangle$. Each X_t is the meaning / situation model $[\![\varphi_t]\!]$ corresponding to the t-th sentence φ_t in the discourse D. Each $X_t \in \Sigma$ is a 150-dim vector, but it is also a single point in the high-dimensional situation-model space Σ. Therefore, the meaning of discourse D can be thought of as a *trajectory* $\bar{X}$, passing through points $X_1, X_2, \ldots, X_T \in \Sigma$ in this specific (temporal) order.

For simplicity, we take narrative temporal progression, i.e., the progression of situation-model time, to precisely mirror the sequence of sentences in discourse: if a sentence φ_2 immediately follows a sentence φ_1 in a discourse, we assume $[\![\varphi_2]\!]$ happens at the time step immediately after the time step at which $[\![\varphi_1]\!]$ holds. This assumption can be relaxed in future work to accommodate a more diverse set of temporal relations between successive sentences in discourse.

The SOM-based vectors $[\![\varphi_1]\!], \ldots, [\![\varphi_T]\!]$ provide the initial sequence of static meanings encoding the non-temporal component of discourse interpretation. But as discussed in Section 1, a full understanding of narrative discourse involves inferential enrichment based on the learned *temporal dynamics* of the world. This temporal dynamics, encoded in the transition matrix W, captures the 'hard' and 'soft' temporal contingencies of the microworld, i.e., how situations typically evolve over time.

Matrix W encodes influences between microfeatures of situation models across *adjacent* time steps. Temporal dependencies across multiple time steps – how events are likely to unfold over longer time horizons – arise exclusively via the composition of influences across single time steps. That is, our temporal dynamics model is a *first-order Markov model* (Golden & Rumelhart 1993; Golden et al. 1994; Frank et al. 2003): we assume that our situation model X_t (what we think is happening at time t) is influenced only by what we think is happening at adjacent times $t \pm 1$, i.e., by the proximate causes in the preceding situation model X_{t-1} and the immediate effects reflected in the subsequent situation model X_{t+1}. The first-order Markov assumption does not rule out temporal dependencies across multiple time steps; it just derives them exclusively by composing single time-step dependencies. This assumption is what enabled us to claim that our learned model of temporal dependencies is fully captured by the $n \times n$ transition matrix W (where $n = 150$).

Given a situation model $[\![\varphi_t]\!]$ representing the state of the world at a time t, we use W to predict the situation model / world state at the next time step $t+1$, and $W^\top$ (the transpose of W) to predict the situation model / world state at the previous time step $t-1$, as shown in (14) and (15) below. This predictive capacity, enabled by the learned matrix W, is crucial for evaluating counterfactuals: we can make predictions about "what happens next" after the counterfactual antecedent, so that we can assess the likelihood of the counterfactual consequent at that subsequent time step.

(14) Predicted situation model at $t+1$: $\quad \text{NEXT}([\![\varphi_t]\!]) := \sigma([\![\varphi_t]\!]W)$

(15) Predicted situation model at $t-1$: $\quad \text{PREV}([\![\varphi_t]\!]) := \sigma([\![\varphi_t]\!]W^\top)$

The result of multiplying the situation-model vector $[\![\varphi_t]\!]$ with the matrix W (or $W^\top$) is a real-valued vector with values potentially outside the $[0,1]$ interval. The non-linear sigmoid transformation σ in (14) and (15) ensures that the predicted activations remain within the valid range $[0,1]$, i.e., that the predicted situation models at $t \pm 1$ are valid points in our situation-model space Σ.[4]

5 Discourse meaning II: updating trajectories to increase temporal coherence

Discourse interpretation, in our account, is a dynamic *process* of seeking temporal coherence. The process begins with the initial 'static' / non-temporal interpretation of the sequence of sentences in the discourse. For a discourse $D = \varphi_1; \varphi_2; \ldots; \varphi_T$, this initial interpretation, which represents what is explicitly stated in D, is the time-indexed trajectory $\bar{X}^{\text{initial}} = \langle X_1^{\text{initial}}, X_2^{\text{initial}}, \ldots, X_T^{\text{initial}} \rangle$. For every time step $t \in 1 \ldots T$ in this sequence / trajectory, $X_t^{\text{initial}} = [\![\varphi_t]\!]$–that is, situation model X_t^{initial} is the 'static' meaning (truth conditions / situation model) of sentence φ_t.

This initial interpretation is just the beginning. The trajectory $\bar{X}^{\text{initial}}$ is then refined and enriched to align it more closely with the learned temporal regularities of the world, encoded in matrix W. We referred to this process, which for our purposes here is discourse interpretation at its core, as *inferential enrichment* based on background temporal world knowledge. Formally, temporal inferential enrichment involves updating / 'moving' the $\bar{X}^{\text{initial}}$ trajectory in closer alignment with the temporal matrix W: we use the temporal dynamics W to enrich the situation model X_t^{initial} at time t with additional inferences we can make about time t based on the immediately preceding situation model X_{t-1}^{initial} and the subsequent situation model X_{t+1}^{initial}. By aligning / enriching each situation model at any given time t in the discourse trajectory $\bar{X}^{\text{initial}}$ with the expectations derived from W, we increase the overall temporal coherence of the discourse interpretation.

The cognitive work of temporal inferential enrichment, formalized as the process of 'pushing' the discourse trajectory through trajectory space Σ^T to achieve better alignment with W, takes some processing time τ. Because this dynamic component of interpretation involves 'moving' meanings in a high-dimensional 'mental representation' space, we are able to inherently connect processing time to semantic interpretation: the greater the 'distance' between the initial trajectory $\bar{X}^{\text{initial}}$ and the final trajectory that is aligned with W, the longer the processing time τ. In other words, interpretation processing time τ is a function of how much inferential enrichment is needed to achieve temporal coherence.

4 A sigmoid function is a continuous, differentiable function that maps real numbers to the interval $(0,1)$, with an S-shaped curve. See Brasoveanu in preparation, Frank et al. 2003 for the specific form of the sigmoid function σ used here.

It is important to clearly distinguish between two notions of time in our semantic framework: τ is continuous time – actual time taken by the human processor to interpret discourses, while t is discrete mental-representation time – situation-model time that is part of the semantic representation constructed by the human processor.

The inferential enrichment / trajectory update process begins with the initial discourse trajectory $\bar{X}^{\text{initial}}$ and unfolds over the processing time τ as follows. On one hand, for each time step t (from 1 to T) in our discourse trajectory with T situation models, we have the current state of the situation model $X_t(\tau)$ at processing time τ. For example, at the beginning of the update process when $\tau = 0$, $X_t(\tau) = X_t(0) = X_t^{\text{initial}}$. And on the other hand, for each time step t, we can compute an *expected situation model* $\mathbb{E}_t(\tau)$, as shown in (16): $\mathbb{E}_t(\tau)$ represents the "best guess" according to W for how events unfolded at discourse time t based exclusively on $X_{t-1}(\tau)$ and $X_{t+1}(\tau)$, i.e., based on $X_t(\tau)$'s immediate temporal neighbors in the trajectory.

$$\text{(16)} \quad \mathbb{E}_t(\tau) = \sigma(\quad X_{t-1}(\tau)W \quad + \quad X_{t+1}(\tau)W^{\top} \quad)$$

To align a discourse trajectory $\bar{X}(\tau)$ with W, each situation model $X_t(\tau)$ in the trajectory is 'pushed' in the direction that $\mathbb{E}_t(\tau)$ points towards. This update occurs in parallel for all situation models $X_t(\tau)$ in the discourse trajectory $\bar{X}(\tau)$, and for all 150 cells / microfeatures in each $X_t(\tau)$. Importantly, this "push" must respect the information explicitly stated in the discourse and captured in the $\bar{X}^{\text{initial}}$ trajectory: the inferential enrichment process can only *conjunctively refine* the stated discourse $\bar{X}^{\text{initial}}$ towards better alignment with world knowledge W, but it cannot contradict this explicitly stated discourse – even if this could result in better alignment with W.

Thus, the temporal-coherence seeking update of the entire discourse trajectory involves the co-evolution of all its constituent situation models during processing time τ in a way that conjunctively refines $\bar{X}^{\text{initial}}$. Formally, the dynamics for the 'movement' / update of the complete discourse trajectory $\bar{X}(\tau) = \langle X_1(\tau), X_2(\tau), \ldots, X_T(\tau)\rangle$ relative to processing time τ is defined by a system of coupled ordinary differential equations (ODEs). This system defines a *vector field* $\mathscr{F}$ in the trajectory space Σ^T:

$$\text{(17)} \quad \mathscr{F}(\bar{X}(\tau), \bar{X}^{\text{initial}}, W) = \left\langle \frac{dX_1(\tau)}{d\tau}, \frac{dX_2(\tau)}{d\tau}, \ldots, \frac{dX_T(\tau)}{d\tau} \right\rangle,$$
where $\bar{X}^{\text{initial}} = \bar{X}(\tau = 0)$ is the initial discourse trajectory consisting of the 'static' / non-temporal SOM-based sentence meanings.

Intuitively, the vector field $\mathscr{F}$ represents a *flow* in trajectory space Σ^T that points any given trajectory $\bar{X}(\tau)$ at any given processing time τ in the direction in which it should evolve to increase its alignment with temporal world knowledge W. Driven by $\mathscr{F}$, the discourse trajectory $\bar{X}(\tau)$ flows towards a configuration that is more aligned with the temporal regularities encoded in W. Flow velocity decreases as the discourse trajectory settles into an equilibrium state / basin of attraction, where

the interpretation $\bar{X}(\tau^\star)$ stabilizes at some final processing time $\tau^\star$ (Brasoveanu in preparation and Frank et al. 2003 provide further details).

The vector field / flow $\mathscr{F}$ induces a gradient-descent-like behavior in a cognitive landscape shaped by both explicit constraints from the stated discourse $\bar{X}^{\text{initial}}$ and implicit constraints from the learned temporal dynamics model W (via $\mathbb{E}_t(\tau)$). The system of ODEs is coupled because the update $\frac{dX_t(\tau)}{d\tau}$ for $X_t(\tau)$ depends on $\mathbb{E}_t(\tau)$, which in turn depends on its neighbors $X_{t-1}(\tau)$ and $X_{t+1}(\tau)$. Symmetrically, updates for these neighbors depend on $X_t(\tau)$. This coupling / co-evolution allows changes to propagate bidirectionally throughout the entire trajectory, as individual situation model vectors influence their neighbors and are influenced by them in return.

As indicated, the $\mathscr{F}$-driven inference process stops when the trajectory $\bar{X}(\tau)$ settles into a stable interpretation $\bar{X}(\tau^\star)$ at time $\tau^\star$. At $\tau^\star$, the interpretation for the current linguistic input is considered complete, and we are ready to interpret the next sentence (or conclude the interpretation). This *stability / end-of-processing time* $\tau^\star$ is controlled by a depth-of-processing parameter $\theta > 0$: $\tau^\star$ is the earliest processing time $\tau \geq 0$ where the overall rate of change in the trajectory (measured by the L1 norm / sum of absolute values of $\mathscr{F}$) falls below a threshold determined by θ:

(18) $\tau^\star$: earliest processing time $\tau \geq 0$ s.t. $\left|\mathscr{F}(\bar{X}(\tau), \bar{X}^{\text{initial}}, W)\right|_1 < \frac{1}{\theta}$

Parameter θ modulates the depth of processing. Large θ values, e.g., $\theta = 10$, correspond to small thresholds $\frac{1}{\theta}$, so the process runs for a longer time $\tau^\star$, leading to deeper processing: inferences fully propagate throughout the trajectory, and when interpreting incrementally, the next sentence is added only when inferential enrichment based on previous sentences is mostly complete. Small θ values, e.g., $\theta = 0.1$, correspond to large thresholds $\frac{1}{\theta}$, which lead to shallower processing: interpretation halts before reaching equilibrium / fully settling into the basin of attraction, and inferences are incomplete when subsequent sentences are added.

We bring all this together in the definition of the *dynamic interpretation function* $\mathbb{I}^{W,\theta}$, which combines the initial static interpretation of the discourse $\bar{X}^{\text{initial}}$ with the dynamic inferential enrichment process driven by W and controlled by θ:

(19) Dynamic Int.: $\mathbb{I}^{W,\theta}(\bar{X}^{\text{initial}}) = \bar{X}(\tau^\star) = \bar{X}^{\text{initial}} + \int_0^{\tau^\star} \mathscr{F}(\bar{X}(\tau), \bar{X}^{\text{initial}}, W)d\tau$

Dynamic interpretation $\mathbb{I}$ solves the system of coupled ODEs defined by $\mathscr{F}$ given the initial trajectory $\bar{X}^{\text{initial}}$, and returns the stabilized trajectory $\bar{X}(\tau^\star)$ at the stability/end time $\tau^\star$ determined by θ. The integral $\int_0^{\tau^\star} \mathscr{F}(\ldots)d\tau$ represents the total accumulated flow or change driven by the vector field $\mathscr{F}$: the total temporal inferential enrichment applied to the initial trajectory $\bar{X}^{\text{initial}}$ over the entire processing time up to $\tau^\star$.

With the dynamic interpretation function $\mathbb{I}$ in hand, we are ready to formally define incremental sentence-by-sentence discourse interpretation. First, recall that the static interpretation function $[\![\cdot]\!]$ (Sections 2-3) assigns initial static meanings

(150-dimensional situation-model vectors from Σ) to propositions φ. Propositions φ are defined recursively in (20). Discourses D can then defined in terms of these 'static' propositions and a *dynamic conjunction* (or sequencing) operator ';' that is used to incrementally extend an existing discourse with a new proposition φ. This means a discourse D is either a single proposition φ, or it is a discourse D' followed by a new proposition φ, also shown in (20) (each boxed for easy identification).

(20) $\boxed{\varphi ::= \text{atomic\_prop} \mid \neg\varphi \mid (\varphi \wedge \psi) \mid (\varphi \vee \psi)}$ $\qquad$ $\boxed{D ::= \varphi \mid D'; \varphi}$

(21) $$[\![D]\!]^{W,\theta} = \begin{cases} \langle [\![\varphi]\!] \rangle \text{ (singleton trajectory w/ static int.)} & \text{if } D := \varphi \\ \mathbb{I}^{W,\theta}\left([\![D']\!]^{W,\theta} \oplus \langle [\![\varphi]\!] \rangle\right) \text{ (non-singleton traj. w/ dyn. int.)} & \text{if } D := D'; \varphi \end{cases}$$

Incremental dynamic discourse interpretation $[\![D]\!]^{W,\theta}$ is then defined as in (21). If the discourse D consists of only a single proposition φ (i.e., $D := \varphi$), its interpretation is a singleton trajectory containing only the static meaning of φ. If the discourse D is formed by extending a previous discourse D' with a new proposition φ, i.e., $D := D'; \varphi$, its interpretation is obtained by applying the dynamic interpretation function $\mathbb{I}^{W,\theta}$ defined in (19) to the trajectory formed by concatenating the *already stabilized* interpretation of D' with the static meaning of the new proposition φ.

Let's illustrate this with the discourse in (3). Set $\theta = 0.3$. The first sentence is interpreted, and we get a singleton trajectory $[\![\text{RAIN}]\!]^{W,0.3} = \langle [\![\text{RAIN}]\!] \rangle$. The second sentence is then interpreted in this context: $[\![\text{RAIN}; (\text{B\_COMP} \wedge \text{J\_COMP})]\!]^{W,0.3} =$ (by (21)) $\mathbb{I}^{W,0.3}([\![\text{RAIN}]\!]^{W,0.3} \oplus \langle [\![\text{B\_COMP} \wedge \text{J\_COMP}]\!] \rangle) = \mathbb{I}^{W,0.3}(\langle [\![\text{RAIN}]\!], [\![\text{B\_COMP} \wedge \text{J\_COMP}]\!] \rangle)$. Thus, we start with the initial 'static' trajectory $\langle X_1^{\text{initial}}, X_2^{\text{initial}} \rangle = \langle [\![\text{RAIN}]\!], [\![\text{B\_COMP} \wedge \text{J\_COMP}]\!] \rangle$ and apply the dynamic interpretation function $\mathbb{I}^{W,0.3}$ to it, which enriches it with temporal inferences based on W. At stability time $\tau^\star$ determined by $\theta = 0.3$, the result is an inferentially-enriched trajectory $\langle X_1^\star, X_2^\star \rangle = \mathbb{I}^{W,0.3}(\langle X_1^{\text{initial}}, X_2^{\text{initial}} \rangle)$ that provides the context for the counterfactual in (4).

6 Counterfactual interpretation as search for temporal coherence

We now turn to the interpretation of counterfactuals (CFs), which leverages the same coherence-seeking semantics framework we developed for factual discourse. While the underlying dynamic mechanism is shared, the meaning and function of counterfactuals are distinct. Unlike indicative sentences or discourses, which are primarily about a current (candidate for an) actual situation model or trajectory, we propose that counterfactuals are fundamentally statements about the learned temporal world knowledge matrix W: they explore the consequences of hypothetical changes to the world *as constrained by the regularities encoded in* W. In intensional semantics terms, conditionals in general are about modal accessibility relations, and counterfactuals specifically are about the historically-structured similarity relation.

In our framework, this similarity relation is emergent, i.e., constructed from the coherence-seeking process constrained by W and the actual context.

Consistent with analyses of past counterfactuals (Ippolito 2013), the counterfactual antecedent is evaluated relative to the past of a relevant point in the actual context: the CF antecedent "If it had been sunny" concerns the state at t_1, which is in the past relative to the relevant actual situation model $X_2^\star$ derived above: the inferentially-enriched situation model at t_2 where Bob and Jane play computer games. The initial representation for the CF antecedent is its static interpretation $[\![\text{SUN}]\!]$. The core of our proposal lies in how the most "similar" counterfactual situation model for the antecedent is determined. Instead of relying on a theoretically primitive similarity metric, counterfactual similarity emerges from temporal-inference enrichment between the CF antecedent $[\![\text{SUN}]\!]$ at t_1 and the actual context $X_2^\star$ at t_2: we apply the dynamic interpretation function $\mathbb{I}$ to the trajectory $\langle [\![\text{SUN}]\!], X_2^\star \rangle$ formed out of the CF antecedent and the actual context state: $\langle [\![\text{SUN}]\!]^\star, X_2^{\star\star} \rangle = \mathbb{I}^{W,0.3}(\langle [\![\text{SUN}]\!], X_2^\star \rangle)$.

The resulting $[\![\text{SUN}]\!]^\star$ situation model is an inferentially-enriched CF antecedent: it is the version of "sunny at t_1" that is maximally temporally coherent with the (subsequent) actual context $X_2^\star$ given our temporal knowledge W. To evaluate the CF consequent, we use W once again and predict what would happen next at t_2 given the CF antecedent $[\![\text{SUN}]\!]^\star$ at t_1, i.e., we compute $\text{NEXT}([\![\text{SUN}]\!]^\star)$. The CF is true iff the conditional probability of its consequent "they would have played outside" given $\text{NEXT}([\![\text{SUN}]\!]^\star)$ exceeds a contextual threshold π. In our microworld, playing outside is translated as the disjunction $[\![\text{SOCCER} \vee \text{HIDE}n\text{SEEK}]\!]$. The word "instead" in the CF signals that this consequent should be compared against the contextual threshold set by the probability of the actual activity that occurred in the original context at t_2 (playing computer games: $[\![\text{B\_COMP} \wedge \text{J\_COMP}]\!]$). So the CF is true iff $\mathbf{P}([\![\text{SOCCER} \vee \text{HIDE}n\text{SEEK}]\!] \mid \text{NEXT}([\![\text{SUN}]\!]^\star)) > \mathbf{P}([\![\text{B\_COMP} \wedge \text{J\_COMP}]\!] \mid \text{NEXT}([\![\text{SUN}]\!]^\star))$. Our model calculates this to be $0.41 > 0.01$, so the CF is true.

7 Summary and remaining challenges

This paper puts forward a semantic theory grounded in a learned model of the world, where sentence meanings are represented as vectors in a high-dimensional space that exhibits meaningful semantic geometry and supports non-temporal and temporal probabilistic inference. Counterfactual interpretation is graded and context-sensitive, and involves an emergent / constructed historically-structured similarity relation. The semantics formally integrates linguistic input and probabilistic world knowledge, and inherently connects interpretation to processing time. Remaining challenges include full sub-sentential compositionality, a flexible mapping between discourse sequencing and temporal progression, and non-past counterfactuals.

References

Brasoveanu, Adrian. in preparation. Reimagining the Semantics of Narratives and Counterfactuals.

Cohn, Neil. 2019. Visual narratives and the mind: Comprehension, cognition, and learning. In Kara D. Federmeier & Diane M. Beck (eds.), *Psychology of Learning and Motivation*, vol. 70, 97–128. London: Academic Press.

Frank, Stefan L., Mathieu Koppen, Leo G. M. Noordman & Wietske Vonk. 2003. Modeling knowledge-based inferences in story comprehension. *Cognitive Science* 27(6). 875–910. doi:10.1207/s15516709cog2706_3.

Frank, Stefan L. & Gabriella Vigliocco. 2011. Sentence comprehension as mental simulation: An information-theoretic perspective. *Information* 2(4). 672–696. doi:10.3390/info2040672.

Golden, Richard M. & David E. Rumelhart. 1993. A parallel distributed processing model of story comprehension and recall. *Discourse Processes* 16(3). 203–237. doi:10.1080/01638539309544839.

Golden, Richard M., David E. Rumelhart, Joseph Strickland & Alice Ting. 1994. Markov random fields for text comprehension. In Daniel S. Levine & M. Aparicio IV (eds.), *Neural Networks for Knowledge Representation and Inference*, 283–309. Hillsdale, NJ: Lawrence Erlbaum Associates. doi:10.4324/9780203763179.

Ippolito, Michela. 2013. *Subjunctive conditionals: a linguistic analysis*. The MIT Press.

Kaufmann, Stefan. 2015. Conditionals, conditional probabilities, and conditionalization. In Henk Zeevat & Hans-Christian Schmitz (eds.), *Bayesian Natural Language Semantics and Pragmatics*, 71–94. Springer.

Khoo, Justin. 2017. Backtracking counterfactuals revisited. *Mind* 126(503). 841–910. doi:10.1093/mind/fzw005.

Kohonen, Teuvo. 1995. *Self-Organizing Maps*, vol. 30 Springer Series in Information Sciences. Berlin, Heidelberg: Springer.

Lassiter, Daniel. 2017. *Graded Modality: Qualitative and Quantitative Perspectives*. Oxford University Press.

Lewis, David K. 1973. *Counterfactuals*. Harvard University Press.

Noordman, Leo G. M. & Wietske Vonk. 2015. Inferences in discourse, psychology of. In *International Encyclopedia of the Social & Behavioral Sciences*, 37–44. Elsevier 2nd edn. doi:10.1016/b978-0-08-097086-8.57012-3.

Prior, Arthur N. 1955. *Formal Logic*. Oxford: Oxford University Press.

Stalnaker, Robert. 1968. A theory of conditionals. In Nicholas Rescher (ed.), *Studies in Logical Theory*, 98–112. Blackwell.

Stein, Nancy L. & Christine G. Glenn. 1979. An analysis of story comprehension

in elementary school children. In Roy O. Freedle (ed.), *New Directions in Discourse Processing*, vol. 2, 53–120. Ablex Publishing.

Thomason, Richmond H. 1970. Indeterminist time and truth-value gaps. *Theoria* 36(3). 264–281. doi:10.1111/j.1755-2567.1970.tb00427.x.

Venhuizen, Noortje J., Petra Hendriks, Matthew W. Crocker & Harm Brouwer. 2022. Distributional formal semantics. *Information and Computation* 287. 104925. doi:10.1016/j.ic.2022.104925.

Zadeh, Lotfi A. 1975. Fuzzy logic and approximate reasoning. *Synthese* 30(3-4). 407–428. doi:10.1007/bf00485052.

Zwaan, Rolf A. 2001. Situation model: Psychological. In *International Encyclopedia of the Social & Behavioral Sciences*, 14137–14141. Elsevier. doi:10.1016/b0-08-043076-7/01550-3.

Zwaan, Rolf A. 2008. Experiential traces and mental simulations in language comprehension. In M. D. Vega, A. M. Glenberg & A. C. Graesser (eds.), *Symbols, Embodiment, and Meaning: Debates on Meaning and Cognition*, 165–180. Oxford University Press. doi:10.1093/acprof:oso/9780199217274.003.0009.

Zwaan, Rolf A. & Gabriel A. Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin* 123(2). 162–185. doi:10.1037/0033-2909.123.2.162.

Adrian Brasoveanu
Department of Linguistics
UC Santa Cruz
1156 High Street
Santa Cruz, CA 95064
abrsvn@ucsc.edu