Reimagining the Semantics of Narratives and Counterfactuals

Adrian Brasoveanu, UC Santa Cruz

UC Davis Linguistics Colloquium—May 18, 2026

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. *(Stein and Glenn, 1979)*

Drawings generated with ChatGPT 4o (May 2025)

Main Idea (I): Discourse Interpretation as Search for Coherence

- Discourse interpretation = construct sequence of situation models (mental representations)
  e.g., the 4 drawings above, showing situations described in discourse unfolding in time

- We use background world knowledge to enrich explicitly stated story with inferences
- Boy silly b/c background world knowledge includes: cakes crumble when squeezed
  t_1: boy puts cake in leaf under arm $\rightarrow$ t_2: cake crumbles into pieces
- Coherence = **alignment with learned model of world's temporal dynamics**

- **Interpretation = coherence-increasing inference process**
  Not *active* goal search, but **constructive process** that takes processing effort / time
- *Compositional* interpretation at the sub-sentential level: problem for another day
  Focus: pervasive 'intrusion' of world knowledge in interpretation construction
- Misleading to call this 'intrusion'—it's the **main work** of discourse interpretation

Main Idea (II): Counterfactual Interpretation as Search for Coherence

In this context, we can say:

If the boy had carried the cake on top of his head, the cake wouldn't have crumbled.

Proposal: Counterfactual Interpretation Leverages the Same Inference Process

- **Same coherence-seeking inference process underlies counterfactual interpretation**
- To interpret CF: construct CF antecedent situation model as similar as possible to actual context by **maximizing temporal coherence of CF antec and actual context**
- CF antec + actual context = "miracle" (incoherent) timeline; coherence-seeking inference smooths it over ("heals" it)

- Follow Thomason (1970); Ippolito (2013); Khoo (2017): historically-structured interpretation for counterfactuals
- But I **construct** the similarity relation via coherence-seeking—formally explicit pragmatics
- Build on learned model of the world: Frank et al. (2003) (Golden et al., 1994; Venhuizen et al., 2022)

Need: Similarity + How Events Typically Unfold + Graded 'Entailment'

What kind of meaning representations do we need to interpret counterfactual above – or this one, which we will be focusing on from now on:

[Context]	It's raining (at t_1). Bob and Jane are playing computer games (at t_2).
[CF]	If it had been sunny (at t_1), they would have played outside instead (at t_2).

We need:

1. **Similarity / Distance between Situation Models**: Construct a situation model where it's sunny at t_1 (instead of rainy), but that's otherwise as **similar** as possible to actual context
2. **Predict How Events Typically Unfold**: If at t_1 we're in a counterfactual sunny situation that's otherwise as similar as possible to the actual world, **predict what would happen next** at time t_2 based on background knowledge of how events typically unfold
3. **Graded 'Entailment'** (Conditional Probability): In that inertially predicted situation model at time t_2, **it's more likely** that Bob and Jane play outside than that they play computer games (actual activity), even though we constructed a counterfactual situation that's as similar as possible to the actual context
   (see Kaufmann 2015; Lassiter 2017 a.o. for probabilistic approaches to modality)

Formal semantics provides little, most action in definition and pragmatics of accessibility reln.s

Bonus Points: Up-front Similarity Predictions + Processing-Time Predictions

4. Want to be **formal, systematic, up-front** about similarity, temporal unfolding, graded ‘entailment’ predictions
   - Minimize informal pragmatic reasoning made precise on a case-by-case basis
   - Set up microworld + learned model *before* analyzing any specific examples

5. Want **a semantic theory that inherently makes processing predictions**
   - Processing time not external to meaning but **intrinsic** to it
   - Semantic incoherence = surprisal ⇒ linking hypothesis emerges from framework
   - Challenges traditional **competence-performance** boundary

6. World knowledge integrated into core interpretation
   - Not “pragmatic intrusion” but the main work of interpretation
   - Challenges traditional **semantics-pragmatics** boundary

Roadmap

Sentences Denote Situation Models – and an Example Microworld

Understanding a sentence = constructing a situation (mental) model of described situation
= understanding conditions under which sentence true ≈ mentally 'simulating' described situation

Situation models: non-linguistic, truth-conditional, **analog** mental representations

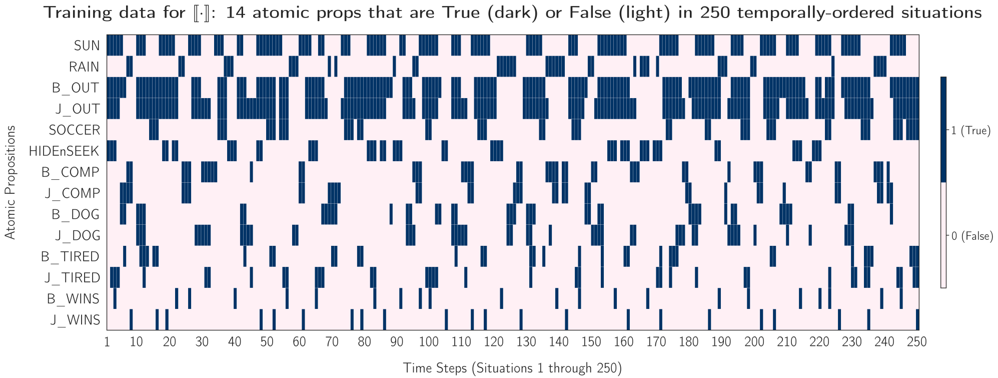
- similar to result of actually perceiving described situation
- disquotational: 'simulate' conditions under which sentence (description of situation) is true
- but more: 'simulation' has analog properties that enable direct inspection for inference

(Zwaan, 1999, 2008; Frank and Vigliocco, 2011; Noordman and Vonk, 2015)

Example microworld: activities and states of two characters Bob and Jane (Frank et al., 2003)
Situation at time t is exhaustively specified by the truth / falsity of 14 atomic propositions:

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-and-seek.	13.	B_WINS	Bob wins.
7.	B_COMP	Bob plays a computer game.	14.	J_WINS	Jane wins.

250 Temporally-Ordered Situations in the Microworld: Training data for ⟦·⟧



Model for Priorian tense logic (Prior 1955 a.o.), but think of it as **training data for ⟦·⟧: sample 'movie' of 250 'time frames' showing how events typically unfold in microworld, which we perceive and learn from**

Columns: complete specification of world @ time t wrt the 14 props/atomic world features = **situations**

Rows: truth-conditions of props, set of world @ times t where prop is 1 (or 0) $\approx$ **situation models**

Why set up a microworld + Two Kinds of World Knowledge

Setting up a microworld in advance means we are **up-front** about our predictions of

> what counts as similar situations, how events typically unfold, what plausibly follows from what
>
> before we see any particular example discourse or counterfactual

World Knowledge: Non-temporal (Within Time Step) and Temporal (Across Time Steps)

Non-temporal constraints – situation models should (approximately) encode these:

- Can't be both sunny and rainy and (probabilistic) it's sunny more often than rainy
- Bob and Jane can only play soccer when they are outside
- They can only play computer games when inside ($\neg$outside)
- They can only perform one activity at a time
- (Probabilistic) They are more likely to be outside when sun shines and inside when it rains

Temporal constraints – drive inferential enrichment during discourse interpretation:

- A game can only be won if it was played in the previous time step
- (Probabilistic) Whoever is tired is less likely to win at the next time step

The Sample of 250 Temporally-Ordered Situations as Training Data for $[\![\cdot]\!]$

- Use 250 temporally-ordered situations as **training data** to learn an approximate mental model of microworld as characterized by 14 atomic props / elementary features of the world

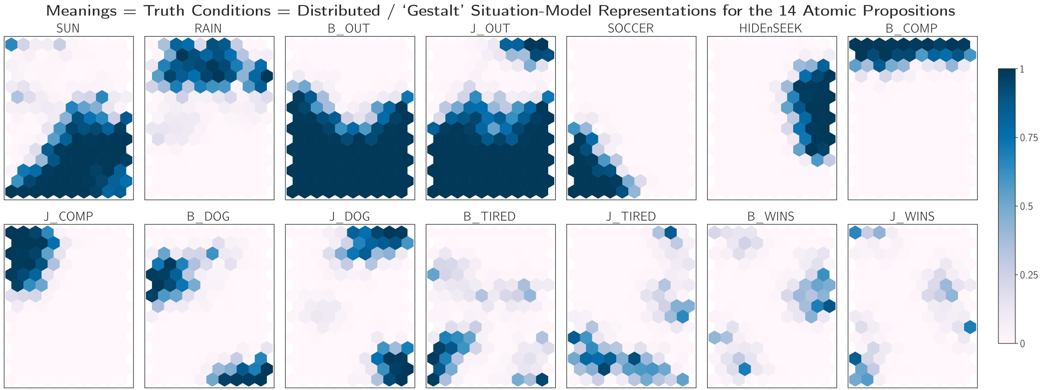
  That is, to **learn the interpretation function** $[\![\cdot]\!]$

- Two components of our learned mental model of the microworld / our learned $[\![\cdot]\!]$:
  - **situation models** encoding (an approx. of) non-temporal 'hard' and 'soft' constraints
  - **temporal dynamics model** encoding (an approx. of) temporal 'hard' and 'soft' constraints

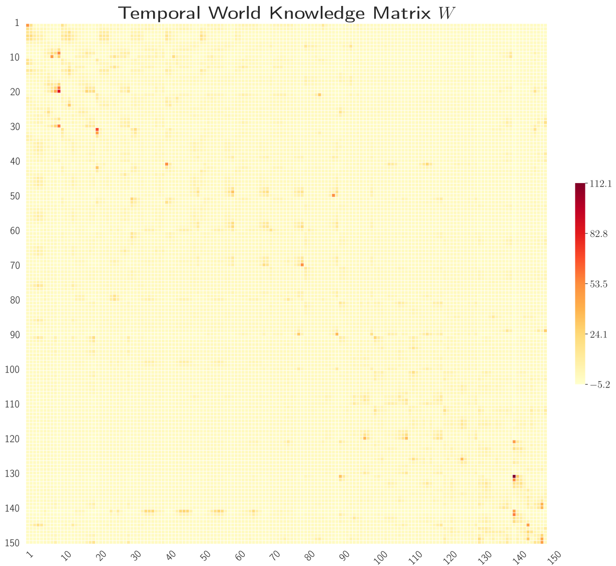
- Various choices for situation / temporal-dynamics models and learning algorithms, go with
  - Self-Organizing Maps (SOMs) for situation models (Kohonen, 1995; Frank et al., 2003)
  - Markov Random Field (MRF) for temporal dynamics (Golden et al., 1994; Frank et al., 2003)

- Specific models and algorithms less important – primarily interested in
  - **analog** nature of learned representations that **directly** supports various operations
  - **operations they support**: similarity, predict future, graded 'entailment', processing time

Learned $[\![\cdot]\!]$ Assigning Truth Conditions = Situation Models to Atomic Props



Darker = higher activation. Patterns encode probabilistic truth-conditions learned from training data.



Entry w_{ij}: association strength between microfeature i at t and microfeature j at $t+1$. Red = positive (facilitatory), yellow = weak/negative (inhibitory).

Sentence Meaning in a Learned Model of the World

- Sentence meanings: situation models formalized as vectors in space $\Sigma = [0,1]^n$ ($n = 150$)
  - each vector: a 15 row $\times$ 10 column map consisting of 150 hexagonal cells
  - each cell has a value in the continuous interval $[0,1]$

- 14 SOM-learned vectors = Lexical meanings for the 14 atomic props: $[\![\text{SUN}]\!], [\![\text{RAIN}]\!], \ldots$
- **Distributed/'Gestalt'** representations of truth conditions (approx. of rows in training data)
  - no 1-1 correspondence between row of 250 truth-values and 150 microfeatures of SOM vector
- Look like language-model embeddings, but fundamentally different: **truly semantic, encode world / non-linguistic patterns**, not word co-occurrence / linguistic patterns

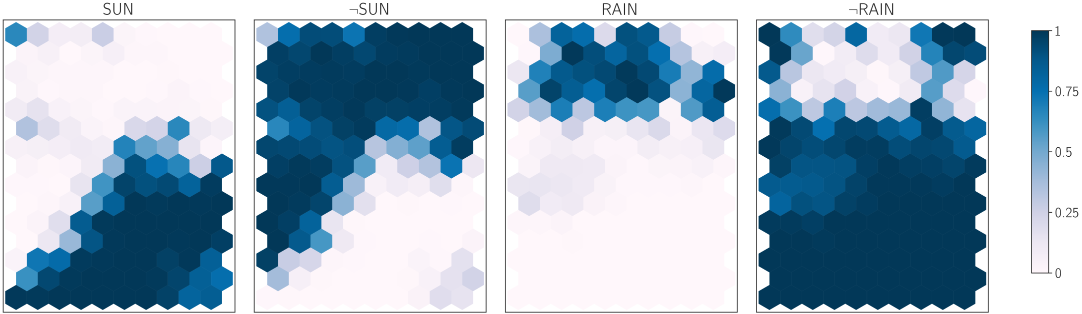
- Have **semantic geometry / analog properties** encoding non-temporal world knowledge:
  - Prior / base probability of occurrence in the microworld $\approx$ average 'height' of map
  - Likelihood of co-occurrence in the microworld $\approx$ overlap / separation of 'peaks'
  - Smaller area that has 'peaks' $\approx$ more specific information, hence lower probability
  - Larger 'peaky' area $\approx$ low info / evidence, supports / compatible with many situations
  - Similarity between meanings $\approx$ distance in situation-model space Σ

Sentence-Level Compositionality Part 1: Negation

With the lexical meanings for atomic props in hand, we recursively assign meanings to any non-atomic prop using vectorial versions of fuzzy logic operations (Zadeh, 1975)

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

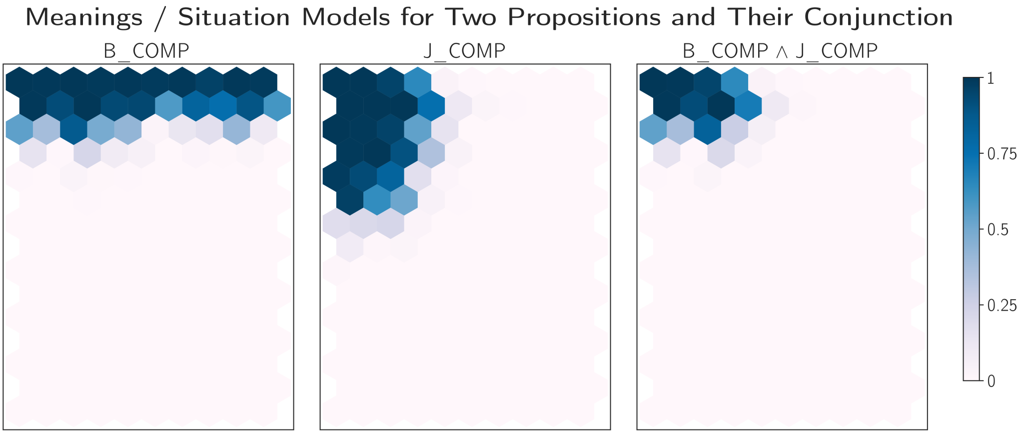
Meanings / Situation Models for Two Propositions and Their Negations



Sentence-Level Compositionality Part 2: Conjunction

Conjunction: $[\![p \land q]\!] = [\![p]\!] \odot [\![q]\!]$ ($\odot$: element-wise / Hadamard product)

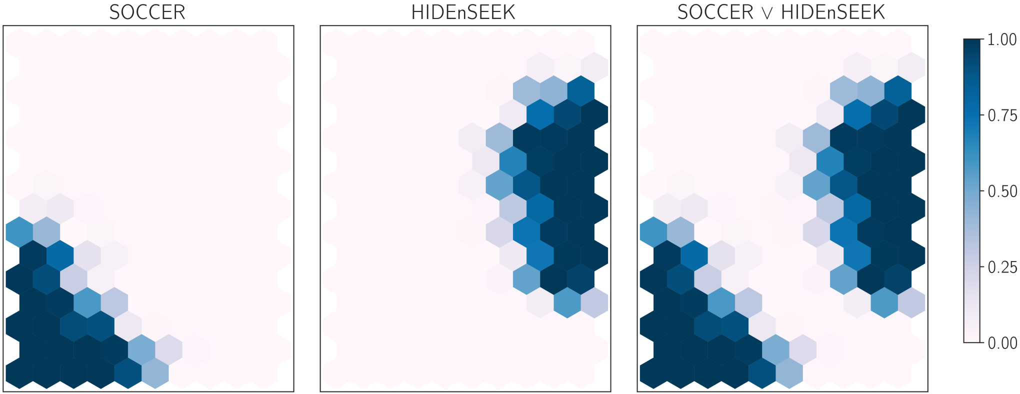
[conjunction could be element-wise min too]



Sentence-Level Compositionality Part 3: Disjunction

Disjunction (by De Morgan): $[\![p \vee q]\!] = [\![\neg(\neg p \wedge \neg q)]\!] = [\![p]\!] + [\![q]\!] - [\![p]\!] \odot [\![q]\!]$

Meanings / Situation Models for Two Propositions and Their Disjunction



Graded 'Entailment' / Conditional Probability

As mentioned, some intuitive analog properties of situation models (SOM vectors) are probabilistic. We can formally define them:

- Prior / base **probability of a prop** p: average activation of its meaning

$$\mathrm{P}(p) = \mathtt{mean}(\llbracket p \rrbracket) = \tfrac{1}{n} \textstyle\sum_i^n \llbracket p \rrbracket_i \qquad = \textstyle\sum_i^n \left(\llbracket p \rrbracket_i \tfrac{1}{n} \right)$$

- Putting together the definitions of probability and conjunction, we define a notion of **graded 'entailment' of a prop p by a prop q = conditional probability of p given q**

$$\mathrm{P}(p \mid q) = \tfrac{\mathrm{P}(p \wedge q)}{\mathrm{P}(q)} \qquad = \frac{\sum_i^n (\llbracket p \rrbracket \odot \llbracket q \rrbracket)_i}{\sum_i^n \llbracket q \rrbracket_i} = \frac{\sum_i^n (\llbracket p \rrbracket_i \cdot \llbracket q \rrbracket_i)}{\sum_i^n \llbracket q \rrbracket_i} = \textstyle\sum_i^n \left(\llbracket p \rrbracket_i \frac{\llbracket q \rrbracket_i}{\sum_i^n \llbracket q \rrbracket_i} \right)$$

Situation-model space Σ supports **graded 'entailment'**, needed for counterfactuals!

Discourse Meaning in a Learned Model of the World

- Meaning of a discourse of length T: **temporally-ordered sequence of situation models** $\bar{X} = \langle X_1, X_2, \ldots, X_T \rangle$, where situation model X_t is meaning of t-th sentence in discourse

- For now, we hard-code narrative temporal progression:
  seq. of sentences in discourse = progression of situation-model time

  (We can generalize this to handle flashbacks, elaborations, explanations, non-past CFs – see Appendix)

- Meaning of discourse: **trajectory in situation-model space** Σ, passing through points $X_1, X_2, \ldots, X_T$ in (temporal) order [situation models are points in Σ]

- But: SOM vectors provide only static meanings

Discourse meanings (temporal progressions) also involve temporal dynamics

The Temporal World Knowledge Matrix W

- **First-order Markov assumption** (simplest assumption): what we think is happening at time t (our situation model at time t, a.k.a. X_t) is directly influenced only by what we think is happening at adjacent times $t \pm 1$

  (Golden and Rumelhart, 1993; Golden et al., 1994; Frank et al., 2003)

- Influences across multiple steps arise exclusively via composing influences across single steps

- **Temporal knowledge: an $n \times n$ matrix W ($n = 150$), estimated from training data**

- W captures all temporal correlations, not only causal connections: what is likely happening at time step t given what's happening at neighboring time steps $t \pm 1$

  Bob tired at t, then less likely to win at $t + 1$
  Bob wins at t, must have been playing some game at $t - 1$
  Rains at t, Bob and Jane likely to move inside at $t + 1$, etc.

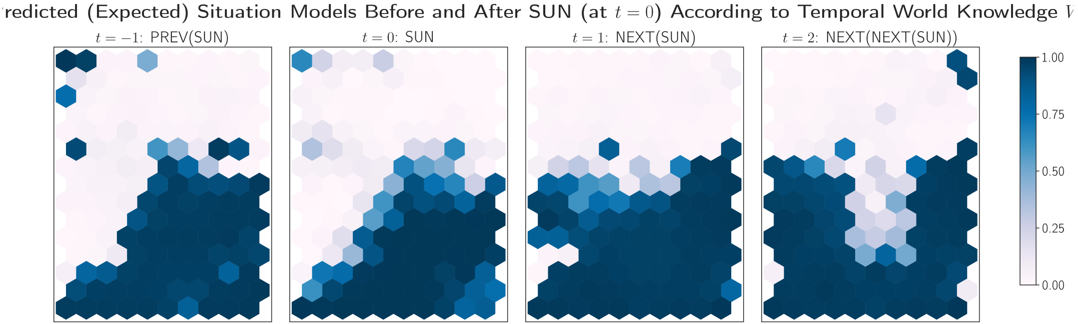
Computing Expected Situation Model at Next / Previous Time Step

- We can now predict next / previous situation model given situation model $[\![p]\!]$ at t:

  Expected at $t+1$: $[\![\texttt{Next}(p)]\!] = \texttt{NEXT}([\![p]\!]) := \sigma([\![p]\!]W)$

  Expected at $t-1$: $[\![\texttt{Previous}(p)]\!] = \texttt{PREV}([\![p]\!]) := \sigma([\![p]\!]W^\top)$ [$W^\top$ is the transpose of W]

  σ: sigmoid function mapping to $(0,1)$; ensures predicted situation models remain valid points in Σ



W supports **predicting how events typically unfold**, needed for counterfactuals!

Discourse Interpretation as Search for Temporal Coherence

Discourse Interpretation 1: Initial Static Meaning [Already Discussed]

Stated discourse meaning = temporally-ordered sequence of situation models (one per sentence)
= time-indexed trajectory in situation-model space $\bar{X} = \langle X_1, X_2, \ldots, X_T \rangle$

Discourse Interpretation 2: Dynamic Update – Increase Alignment with Knowledge W

Interpretation = Inference Process that Moves Initial Meaning in Closer Alignment with W

- Use background temporal world knowledge W to enrich stated discourse with inferences
- Inferential enrichment increases temporal coherence, i.e., aligns discourse trajectory with W
- Work of interpretation / inference: pushing discourse trajectory in situation-model space Σ
- This work / cognitive effort takes processing time τ

τ is continuous time – actual processing / reaction time
t is discrete time – narrative / situation-model time

Search for Coherence = Increase Alignment with World Knowledge W

- Initial discourse trajectory (static interpretation of stated discourse): $\bar{X} = \langle X_1, X_2, \ldots, X_T \rangle$

- For each $t = 1 \ldots T$, compute expected situation model at t given situation models at $t \pm 1$:
  $\mathbb{E}_t = \sigma(X_{t-1} W + X_{t+1} W^\top)$

  $\mathbb{E}_t$: best guess according to W for how events unfolded at t if all we knew was $X_{t\pm1}$

- **To align discourse trajectory $\bar{X}$ with W, push each X_t in direction $\mathbb{E}_t$ points towards**

- Inference process / dynamic interpretation happens incrementally, one sentence at a time:
  - Sentence 1 is statically interpreted: denotes situation model X_1
  - Sentence 2 is statically interpreted: denotes situation model X_2
  - Dynamic update: align partial trajectory $\langle X_1, X_2 \rangle$ with W
  - Sentence 3 is statically interpreted: denotes X_3; then dynamically update $\langle X_1, X_2, X_3 \rangle$
  - Continue with the static + dynamic interpretation of next sentence, until final sentence T

ODEs for the Dynamics of Trajectory Update $\Longrightarrow$ Similarity

Dynamics for complete discourse trajectory: **vector field** $\mathcal{F}$ defined by system of **coupled** ODEs

$$\boxed{\mathcal{F}(\bar{X}(\tau), \bar{X}^{\text{initial}}, W)} \quad = \quad \left\langle \frac{dX_1(\tau)}{d\tau}, \frac{dX_2(\tau)}{d\tau}, \ldots, \frac{dX_T(\tau)}{d\tau} \right\rangle$$

- $\mathcal{F}$: **flow** in trajectory space Σ^T, points trajectory toward alignment with W
- Flow slows as system approaches equilibrium; stabilizes at time $\tau^\star$ (controlled by θ)
  (θ = depth-of-processing parameter controlling how long inference runs before stabilizing)
- An **Anchor term** in each ODE (not shown) ensures activations stay in $[0, X^{\text{initial}}]$:
  - Ceiling X^{initial}: can't exceed what was explicitly stated
  - Floor 0: incoherent microfeatures decay toward zero
  - Interpretation can only **conjunctively refine**, never contradict input

ODEs provide **similarity** for CFs: sculpt CF antec to be maximally coherent with actual context, while Anchor ensures CF antec remains true

Dynamic Conjunction and Incremental Dynamic Discourse Interpretation

1. Interpretation function $[\![\cdot]\!]$ assigns initial static meanings $\bar{X}^{\text{initial}}$ to props p in discourse
2. Given initial static trajectory $\bar{X}^{\text{initial}}$ (explicitly stated), the final temporal-inference enriched interpretation is provided by the **dynamic interpretation function** $\mathcal{I}^{W,\theta}(\bar{X}^{\text{initial}})$:

$$\boxed{\mathcal{I}^{W,\theta}(\bar{X}^{\text{initial}}) \quad = \quad \bar{X}(\tau^\star)} \quad = \quad \bar{X}^{\text{initial}} \quad + \quad \int_0^{\tau^\star} \mathcal{F}(\bar{X}(\tau), \bar{X}^{\text{initial}}, W)d\tau$$

Discourses D defined in terms of a **dynamic conjunction** (sequencing) operator ;

$D := p \mid D; p$ We use ; to incrementally extend discourse with one static prop p

Incremental Dynamic Discourse Interpretation $[\![D]\!]^{W,\theta}$ for a discourse $D \ := \ D'; p$

$$\boxed{[\![D]\!]^{W,\theta} = \mathcal{I}^{W,\theta}([\![D']\!]^{W,\theta} \oplus \langle [\![p]\!] \rangle)}$$

$[\![D']\!]^{W,\theta} \oplus \langle [\![p]\!] \rangle$: extension of trajectory $[\![D']\!]^{W,\theta}$ with $[\![p]\!]$ $\oplus$: sequence concatenation

$[\![D']\!]^{W,\theta}$: final, already stabilized interpretation of prefix discourse D'

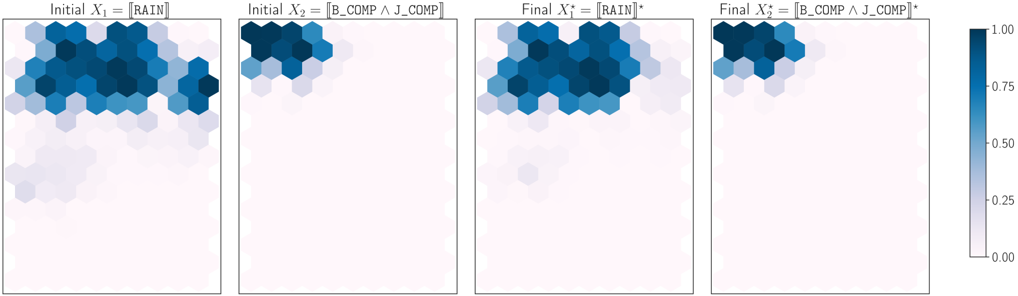
Example: It's raining. Bob and Jane are playing computer games.

Example Discourse: The Context for Our Counterfactual $\theta = 0.3$

It's raining (at $t = 1$). Bob and Jane are playing computer games (at $t = 2$).

$$[\![\texttt{RAIN}; (\texttt{B\_COMP} \land \texttt{J\_COMP})]\!]^{W,0.3} \;=\; \mathcal{I}^{W,0.3}(\langle [\![\texttt{RAIN}]\!], [\![\texttt{B\_COMP} \land \texttt{J\_COMP}]\!] \rangle) \;=\; \langle X_1^\star, X_2^\star \rangle$$

Two-Sentence Dyn. Conj.: Initial Trajectory $\langle X_1, X_2 \rangle$, and Final Trajectory $\langle X_1^\star, X_2^\star \rangle$ after Dyn. Interpretation



CF Interpretation as Search for Temporal Coherence

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

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

- Counterfactuals leverage the same coherence-seeking inferential process as discourses

  But their meaning is of a different kind:

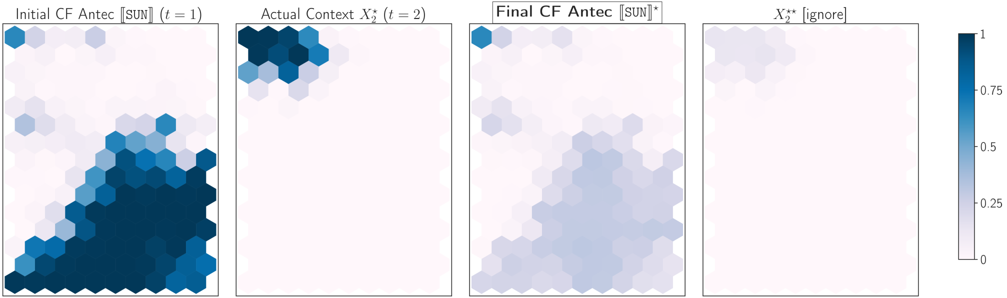
  Ind. sentences/discourses: about current (candidate for) actual situation model/trajectory
  Counterfactuals: statements about the temporal world knowledge matrix W
  In intensional semantics terms:
  - Conditionals in general: about modal accessibility relations
  - Counterfactuals specifically: about the historically-structured similarity relation
- So we will assemble their interpretation differently:

  Context for CF Interpretation: $\mathcal{I}^{W,0.3}(\langle [\![\mathtt{RAIN}]\!], [\![\mathtt{B\_COMP} \wedge \mathtt{J\_COMP}]\!] \rangle) = \langle X_1^\star, X_2^\star \rangle$
  We set CF antecedent in the past of actual context $X_2^\star$ (Ippolito, 2013), then **2 steps**:
  1. Update situation model for CF antecedent to be as similar as possible to actual context $X_2^\star$
  2. Evaluate CF consequent relative to time step after the updated CF antecedent

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

1. Update situation model for CF antecedent to be as similar as possible to actual context $X_2^\star$
   - REPLACE operation: swap actual $X_1^\star$ for CF antec $[\![\textsc{sun}]\!]$; then coherence-seeking "heals" the hybrid timeline
   - $[\![\textsc{sun}]\!]^\star$: CF antec maximally coherent with actual context $(\langle [\![\textsc{sun}]\!]^\star, X_2^{\star\star}\rangle = \mathcal{I}^{W,0.3}(\langle [\![\textsc{sun}]\!], X_2^\star\rangle))$

CF Antec Situation Model Before and After Similarity Alignment with Actual Context $X_2^\star$ (and W)



CF Interpretation as Search for Temporal Coherence: Step 2

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

2. Evaluate CF consequent relative to time step after CF antecedent $\mathtt{NEXT}([\![\mathtt{SUN}]\!]^\star)$: had it been sunny (at t_1), what would we expect to happen next (at t_2) according to W?

CF true iff cond. prob. of CF conseq given $\mathtt{NEXT}([\![\mathtt{SUN}]\!]^\star)$ $>$ contextual threshold π

$$\mathtt{P}([\![\mathtt{SOCCER} \vee \mathtt{HIDEnSEEK}]\!] \mid \mathtt{NEXT}([\![\mathtt{SUN}]\!]^\star)) > \mathtt{P}([\![\mathtt{B\_COMP} \wedge \mathtt{J\_COMP}]\!] \mid \mathtt{NEXT}([\![\mathtt{SUN}]\!]^\star))$$

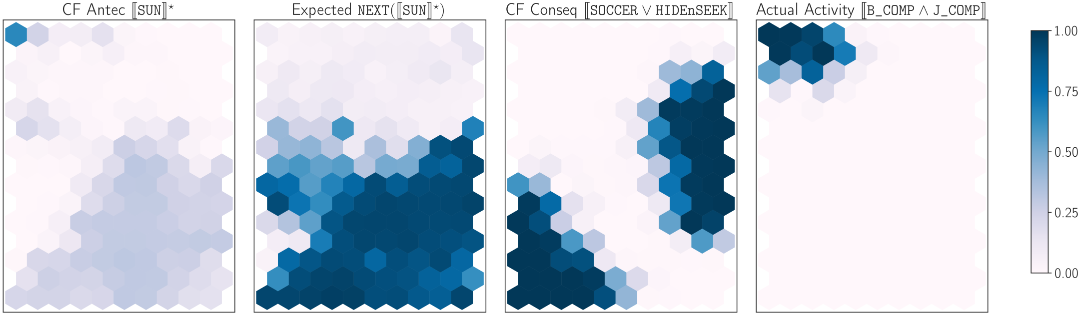
- In our microworld, playing outside is translated as the disjunction $\mathtt{SOCCER} \vee \mathtt{HIDEnSEEK}$
- 'Instead' sets context threshold π to cond. prob. of actual activity given $\mathtt{NEXT}([\![\mathtt{SUN}]\!]^\star)$

CF Interpretation as Search for Temporal Coherence: Step 2 ctd

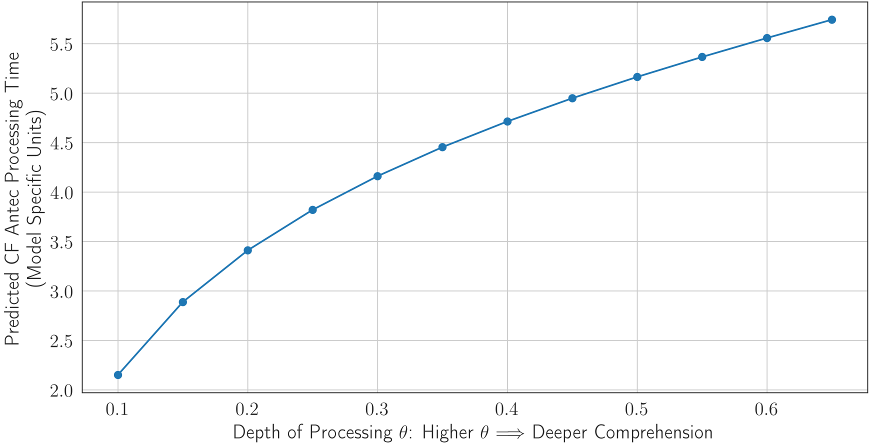
[CF] If it had been sunny (at t_1), they would have played outside instead (at t_2).

$\mathtt{P}(\llbracket\mathtt{SOCCER} \vee \mathtt{HIDEnSEEK}\rrbracket \mid \mathtt{NEXT}(\llbracket\mathtt{SUN}\rrbracket^{\star})) = 0.41$ $\mathtt{P}(\llbracket\mathtt{B\_COMP} \wedge \mathtt{J\_COMP}\rrbracket \mid \mathtt{NEXT}(\llbracket\mathtt{SUN}\rrbracket^{\star})) = 0.01$

Final CF Antec $\llbracket\mathtt{SUN}\rrbracket^{\star}$, Expected $\mathtt{NEXT}(\llbracket\mathtt{SUN}\rrbracket^{\star})$, CF Conseq $\llbracket\mathtt{SOCCER} \vee \mathtt{HIDEnSEEK}\rrbracket$, and Actual Activity $\llbracket\mathtt{B\_COMP} \wedge \mathtt{J\_COMP}\rrbracket$



Processing Similarity: Processing Time for CF Antec $[\![\texttt{SUN}]\!]^{\star}$ as Func. of θ



Processing Similarity ctd: High Incoherence = High Energy

Previous plot: **stopping time of ODE-driven similarity computation process** for CF antecedent:

$$\bar{X}^{\text{initial}}_{\text{CF}} = \langle [\![\mathtt{SUN}]\!], X_2^\star \rangle \quad \xrightarrow{\mathcal{I}^{W,\theta}} \quad \bar{X}^\star_{\text{CF}} = \langle [\![\mathtt{SUN}]\!]^\star, X_2^{\star\star} \rangle$$

- At $\tau = 0$, the CF computation starts with a "miracle" trajectory $\bar{X}^{\text{initial}}_{\text{CF}} = \langle [\![\mathtt{SUN}]\!], X_2^\star \rangle$ that has **high incoherence / high semantic tension / high energy**

  we force sunny weather into past while retaining actual context (Bob & Jane play computer games)

- We quantify **incoherence / tension / energy** for a trajectory $\bar{X}$ as: $\boxed{H(\bar{X}) = -\sum_{t=1}^{T-1} X_t W X_{t+1}^\top}$

- Low energy: neighboring situation models align with W.

  Prediction for next situation $X_t W$ aligns / has high dot product with actual next situation X_{t+1} $\rightarrow$ energy / incoherence $-X_t W X_{t+1}^\top$ is low (really negative).

- High energy: neighboring situation models clash with temporal regularities encoded in W.

  Prediction for next situation $X_t W$ doesn't align / has low dot product with actual next sit. X_{t+1} $\rightarrow$ energy / incoherence $-X_t W X_{t+1}^\top$ is higher (less negative / positive).

Processing Time ctd: High Initial Energy $\approx$ Long Processing Time

- The ODE flow $\mathcal{F}$ pushes the trajectory $\bar{X}$ toward lower H.
- The higher the initial energy $H(\bar{X}^{\text{initial}})$, the longer it takes for the ODE to dissipate it and stabilize at a low-energy trajectory $\bar{X}^{\star}$.

  (θ also controls how long ODE runs before stability: higher θ = deeper comprehension = longer time)
- This is just a form of the surprisal-processing time hypothesis (Hale, 2001; Levy, 2008), but **for semantic interpretation**—the same energy landscape defines trajectory probabilities:

$$P(\bar{X}) = \frac{1}{Z}\exp(-H(\bar{X})) \text{ [Gibbs distribution]} \implies -\ln P(\bar{X}) \text{ [surprisal]} = H(\bar{X}) + \text{constant}$$

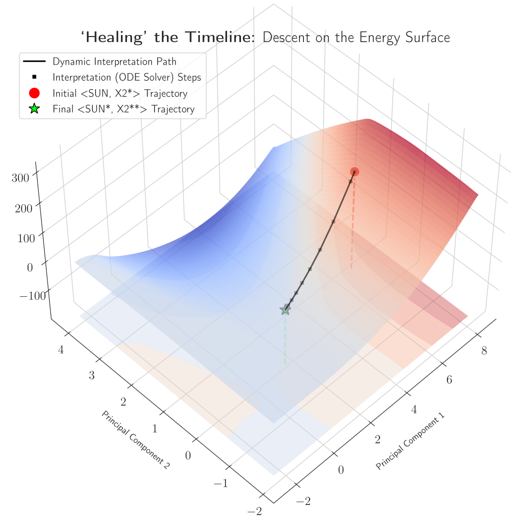
  Energy is surprisal (up to an additive constant).

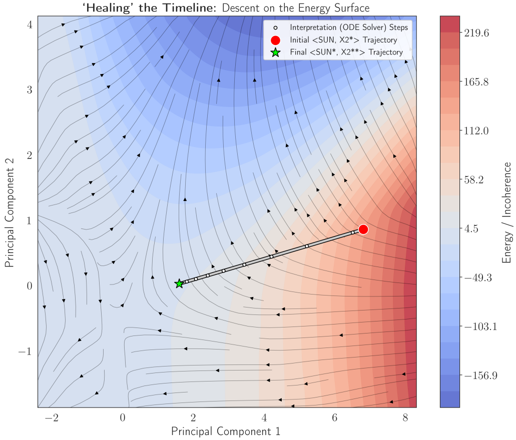
Processing prediction

The integration time $\tau^{\star}$ needed for the ODE to stop correlates with the amount of energy / surprisal dissipated during the interpretation ('relaxation') process:

$$\text{processing time / difficulty} \;\propto\; H(\bar{X}^{\text{initial}}) - H(\bar{X}^{\star})$$

‘Healing’ CF Timeline (Computing Similarity): Descent on Energy Surface





Interpretation as Retrieval from Content-Addressable Memory

DSS architecture—continuous states, sigmoid activation, energy minimization—places it in the class of **Continuous Hopfield Networks (CHNs)** (Hopfield, 1984)

Interpretation = Pattern Completion in Content-Addressable Memory

- **Storage**: "Memories" are not snapshots of specific episodes but *temporal laws*—the "common-sense physics" of how histories typically unfold, encoded in W
- **Cue**: The initial trajectory $\bar{X}^{\text{initial}}$ is a partial/noisy pattern lacking full causal connectivity; for CFs, a deliberately distorted "miracle" timeline
- **Retrieval**: The ODE dynamics $\mathcal{I}^{W,\theta}$ = relaxation into nearest attractor basin = *constructive* pattern completion according to learned physics

Key distinction: The network doesn't memorize histories—it memorizes *how histories work*. Retrieval: constructing a coherent timeline, not recalling a stored one.

Two Perspectives on the Same Process

Semantic perspective:

- Interpretation = search for coherence with learned world knowledge
- Counterfactual similarity = constructive, not primitive
- Processing time reflects inferential work

Memory perspective:

- Interpretation = pattern completion from a partial cue
- CF antecedent + actual context = memory cue
- Retrieval = relaxation to nearest basin

Why does the CHN framing matter?

Not competing theories, two perspectives on the same phenomenon. CHN framing buys us:

- Connection to established neural architecture lit., e.g., modern CHNs = transformers
- **An answer to "what does it mean to understand a discourse?":**

  (Re)constructing from memory how the world works, given a partial cue (what is literally said)
  Back to the beginning of the talk: the mental simulation view of meaning

Summary and Conclusion

Distributed Situation-model Semantics (DSS): A semantic theory rebuilt from scratch with world knowledge inferences and processing predictions as intrinsic features

- Situation models: probabilistic representations of truth conditions with meaningful geometry
- Temporal dynamics W: learned "common-sense physics"
- World knowledge (SOM, W) core part of interpretation, not "intrusion"

- Interpretation = seeking coherence with W
- CF similarity: emergent, not primitive
- Coherence = Energy = Surprisal $\Rightarrow$ processing time intrinsic to interpretation, no separate linking theory needed
- Interpretation = Constructive memory retrieval (Continuous Hopfield Network)

Acknowledgments

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 at the UCLA Syntax & Semantics Seminar (May 7, 2025), UCSC S-Circle (May 14, 2025), SALT 35 (May 20-22, 2025), UCSC Psychology Colloquium (Oct. 10, 2025) and the SALT 35 anonymous reviewers for all their comments 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 the Frank et al. (2003) paper, which was the starting point for the Python implementation of the present account.

Selected References I

Asher, N. (2013). Implicatures and discourse structure. *Lingua*, 132:13–28.

Frank, S. L., Koppen, M., Noordman, L. G. M., and Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science*, 27(6):875–910.

Frank, S. L. and Vigliocco, G. (2011). Sentence comprehension and semantic systematicity: The role of dynamically updated, probabilistic situation models. *Discourse Processes*, 48(4):239–263.

Golden, R. M. and Rumelhart, D. E. (1993). A parallel distributed processing model of story comprehension and recall. *Discourse Processes*, 16(3):203–237.

Golden, R. M., Rumelhart, D. E., Strickland, J., and Ting, A. (1994). Markov random fields for text comprehension. In Levine, D. S. and Aparicio IV, M., editors, *Neural Networks for Knowledge Representation and Inference*, pages 283–309. Lawrence Erlbaum Associates, Hillsdale, NJ.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pages 159–166, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092.

Ippolito, M. (2013). *Subjunctive conditionals: a linguistic analysis*. The MIT Press.

Selected References II

Kaufmann, S. (2015). Conditionals, conditional probabilities, and conditionalization. In Zeevat, H. and Schmitz, H.-C., editors, *Bayesian Natural Language Semantics and Pragmatics*, pages 71–94. Springer.

Khoo, J. (2017). Backtracking counterfactuals revisited. *Mind*, 126(503):841–910.

Kohonen, T. (1995). *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg.

Lassiter, D. (2017). *Graded Modality: Qualitative and Quantitative Perspectives*. Oxford University Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Noordman, L. G. and Vonk, W. (2015). Knowledge-based inferences in text comprehension. *Discourse Processes*, 52(7):523–544.

Stein, N. L. and Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. *New directions in discourse processing*, 2:53–120.

Thomason, R. H. (1970). Indeterminist time and truth-value gaps. *Theoria*, 36(3):264–281.

Venhuizen, N. J., Hendriks, P., Crocker, M. W., and Brouwer, H. (2022). Distributional formal semantics. *Information and Computation*, 287:104925.

Zadeh, L. A. (1975). Fuzzy logic and approximate reasoning. *Synthese*, 30(3-4):407–428.

Selected References III

Zwaan, R. A. (1999). Situation models: The mental leap into imagined worlds. *Current Directions in Psychological Science*, 8(1):15–18.

Zwaan, R. A. (2008). Experiential traces and mental simulations in language comprehension. In Vega, M. D., Glenberg, A. M., and Graesser, A. C., editors, *Symbols, Embodiment, and Meaning: Debates on Meaning and Cognition*, pages 165–180. Oxford University Press.

Appendix: Defining the ODE-based Interpretation Process

Let $\bar{X}(\tau) = \langle X_1(\tau), \ldots, X_T(\tau) \rangle$ be a discourse trajectory over processing time τ, with $X_t(\tau) \in [0,1]^n$ ($n = 150$) and $\bar{X}(0) = \bar{X}^{\text{initial}}$.

Expected state at narrative time t:

$$\mathbb{E}_t(\tau) = \sigma_{\text{DSS}}\left(\underbrace{X_{t-1}(\tau) W}_{\text{prediction from past}} + \underbrace{X_{t+1}(\tau) W^\top}_{\text{diagnosis from future}} \right)$$

with boundary conditions $X_0(\tau) = X_{T+1}(\tau) = \mathbf{0}$. The DSS sigmoid is applied componentwise:

$$\sigma_{\text{DSS}}(z) = \frac{1}{1 - e^{-z}} - \frac{1}{z}, \qquad \sigma_{\text{DSS}}(0) = \frac{1}{2} \quad \text{(removable singularity)}.$$

ODE for microfeature (SOM cell) i in X_t:

$$\frac{dX_{ti}(\tau)}{d\tau} = \left(\mathbb{E}_{ti}(\tau) - \frac{1}{2} \right) \begin{cases} X_{ti}^{\text{initial}} - X_{ti}(\tau), & \text{if } \mathbb{E}_{ti}(\tau) > \frac{1}{2} \quad \text{(growth toward ceiling)} \\ X_{ti}(\tau), & \text{if } \mathbb{E}_{ti}(\tau) \leq \frac{1}{2} \quad \text{(decay toward floor)} \end{cases}$$

First term: direction/strength of the push from W.

Second term: ‘anchor’ to what was literally said (X^{initial}: ceiling fixed by the linguistic input; 0: floor.)

Defining the ODE-based Interpretation Process ctd

Vector form for all microfeatures at time t:

$$\frac{dX_t(\tau)}{d\tau} = \left(\mathbb{E}_t(\tau) - \frac{\mathbf{1}}{2}\right) \odot \left[\left(X_t^{\text{initial}} - X_t(\tau)\right) \odot M_t^{+}(\tau) + X_t(\tau) \odot M_t^{-}(\tau)\right]$$

where $\odot$ is element-wise multiplication, $M_{ti}^{+}(\tau) = 1$ iff $\mathbb{E}_{ti}(\tau) > \frac{1}{2}$, and $M_t^{-}(\tau) = \mathbf{1} - M_t^{+}(\tau)$.

Flow field for the full trajectory:

$$\mathcal{F}(\bar{X}(\tau), \bar{X}^{\text{initial}}, W) = \left\langle \frac{dX_1(\tau)}{d\tau}, \frac{dX_2(\tau)}{d\tau}, \ldots, \frac{dX_T(\tau)}{d\tau} \right\rangle$$

Stability time determined by depth-of-processing θ:

$$\tau^\star = \text{earliest } \tau \geq 0 \text{ such that } \left\|\mathcal{F}(\bar{X}(\tau), \bar{X}^{\text{initial}}, W)\right\|_1 = \sum_{t=1}^{T} \sum_{i=1}^{n} \left| \frac{dX_{ti}(\tau)}{d\tau} \right| < \frac{1}{\theta}$$

Appendix: Richer Temporal Structures Beyond Narrative Progression

So far, we hardcoded sentence order = temporal order (narrative progression)

But natural language discourses have richer temporal structures:

1. **Flashback**: Bob won the game. He had practiced all morning.
   Second sentence describes event *prior to* first—past perfect signals backward shift
2. **Elaboration**: Bob and Jane played outside. They played soccer.
   Second sentence provides more specific info about *same* event, not a subsequent one
3. **Explanation**: The cake crumbled into pieces. The boy had squeezed it under his arm.
   Second sentence describes prior event that *caused* the first

Under narrative progression, all three would place second event *after* first

Key insight: The coherence-seeking ODE mechanism is independent of *how* the trajectory is constructed. What varies is the **structural attachment** of new situation models.

Reference Time (Focus) and Event Time (Target)

To track where new information attaches, we maintain:

- **Focus index** f: the currently "active" temporal position
  - Analogous to Reichenbach's *reference time*
  - The temporal vantage point relative to which new events are located

- **Target index** k: where in the trajectory the new situation model belongs
  - Analogous to Reichenbach's *event time*
  - Determined by linguistic cues: tense/aspect, adverbials, discourse connectives

Relation	**Indices**	**Examples**
Narrative progression	$k = f + 1$	simple past after simple past
Flashback / Explanation	$k < f$	past perfect
Elaboration / Simultaneity	$k = f$	progressive, statives

After processing, focus typically updates: $f \leftarrow k$

Elementary Trajectory Operations

Three elementary operations for building/modifying trajectories:

$\textsc{insert\_}\mathbf{1}(\bar{X}, k)$: Extend timeline (no content yet)

$\langle X_1, \ldots, X_{k-1}, \mathbf{1}, X_k, \ldots, X_T \rangle$ $\quad f \leftarrow k$

Inserts unconstrained **1** (all ones) at position k; shifts subsequent elements; updates focus

$\textsc{conjoin}(\bar{X}, [\![\varphi]\!], k)$: Add linguistic constraints at position k

$\langle X_1, \ldots, X_{k-1}, X_k \odot [\![\varphi]\!], X_{k+1}, \ldots, X_T \rangle$ $\quad$ focus unchanged

Hadamard product adds constraints from $[\![\varphi]\!]$ to existing situation model at k

$\textsc{replace}(\bar{X}, [\![\varphi]\!], k)$: Overwrite content (CFs, corrections)

$\langle X_1, \ldots, X_{k-1}, [\![\varphi]\!], X_{k+1}, \ldots, X_T \rangle$ $\quad f \leftarrow k$ (context-dep.)

Substitutes $[\![\varphi]\!]$ for whatever was at position k; used for counterfactuals, corrections

Derived Operations: INSERT and APPEND

The elementary operations compose to give us familiar discourse moves:

$\textsc{insert}(\bar{X}, [\![\varphi]\!], k)$ = extend timeline + add content

$$\textsc{insert}(\bar{X}, [\![\varphi]\!], k) \;:=\; \textsc{conjoin}(\textsc{insert\_1}(\bar{X}, k),\ [\![\varphi]\!],\ k)$$

Step 1: INSERT_1 creates a new unconstrained position at k
Step 2: CONJOIN fills that position with the content $[\![\varphi]\!]$

$\textsc{append}(\bar{X}, [\![\varphi]\!])$ = INSERT at end of trajectory

$$\textsc{append}(\bar{X}, [\![\varphi]\!]) \;:=\; \textsc{insert}(\bar{X}, [\![\varphi]\!], T + 1)$$

This is narrative progression: add new event after everything so far

Payoff: Flashbacks use INSERT with $k < f$; elaborations use CONJOIN at $k = f$; counterfactuals use REPLACE; narrative progression uses APPEND

Generalized Dynamic Interpretation

The dynamic interpretation function $\mathcal{I}^{W,\theta}$ operates on whatever trajectory it receives.

Only the **construction** of that trajectory varies across temporal relations:

Generalized Interpretation Rule

$$\boxed{[\![D';\varphi]\!]^{W,\theta} \;=\; \mathcal{I}^{W,\theta}\left(\,\textsc{op}\left([\![D']\!]^{W,\theta},[\![\varphi]\!],k\right)\,\right)}$$

- $\textsc{op} \in \{\textsc{insert}, \textsc{conjoin}, \textsc{replace}, \ldots\}$: determined by temporal/rhetorical relation
- k: target position, determined by linguistic cues (tense, aspect, connectives)
- $[\![D']\!]^{W,\theta}$: already-stabilized interpretation of preceding discourse

The original rule is recovered as the special case $\textsc{op} = \textsc{append}$, $k = T + 1$:

$$[\![D';\varphi]\!]^{W,\theta} = \mathcal{I}^{W,\theta}([\![D']\!]^{W,\theta} \oplus \langle[\![\varphi]\!]\rangle)$$

Example: Explanation as $\textsc{insert}$

The cake crumbled into pieces (at t_2). The boy had squeezed it under his arm (at t_1).

Step 1: Interpret first sentence as singleton trajectory

$$[\![\text{“The cake crumbled”}]\!]^{W,\theta} = \langle [\![\textsc{crumbled}]\!] \rangle \qquad f = 1$$

Step 2: Past perfect signals $k < f$. Apply $\textsc{insert}$ at $k = 1$:

$$\textsc{insert}(\langle [\![\textsc{crumbled}]\!] \rangle, [\![\textsc{squeezed}]\!], 1) = \langle [\![\textsc{squeezed}]\!], [\![\textsc{crumbled}]\!] \rangle \qquad f \leftarrow 1$$

Now trajectory has correct temporal order: squeezing at t_1, crumbling at t_2

Step 3: Apply dynamic interpretation $\mathcal{I}^{W,\theta}$:

$$\mathcal{I}^{W,\theta}(\langle [\![\textsc{squeezed}]\!], [\![\textsc{crumbled}]\!] \rangle) = \langle X_1^\star, X_2^\star \rangle$$

Coherence-seeking enriches both with the **causal connection** via W

(Under narrative-progression, we'd incorrectly get $\langle [\![\textsc{crumbled}]\!], [\![\textsc{squeezed}]\!] \rangle$—effect before cause)

Counterfactuals as $\textsc{replace}$: A Unified View

The CF similarity computation from earlier is exactly $\textsc{replace}$:

CF Antecedent Interpretation = $\textsc{replace}$ + Coherence-Seeking

Given actual context trajectory $\langle X_1^\star, X_2^\star \rangle$, the CF antecedent "If it had been sunny..." triggers $\textsc{replace}$ at $k = 1$:

$$\textsc{replace}(\langle X_1^\star, X_2^\star \rangle, [\![\textsc{sun}]\!], 1) = \langle [\![\textsc{sun}]\!], X_2^\star \rangle \qquad \text{"miracle" timeline}$$

Then coherence-seeking "heals" this hybrid trajectory:

$$\mathcal{I}^{W,\theta}(\langle [\![\textsc{sun}]\!], X_2^\star \rangle) = \langle [\![\textsc{sun}]\!]^\star, X_2^{\star\star} \rangle$$

The **Anchor term** in the ODE ensures $[\![\textsc{sun}]\!]^\star$ still entails $[\![\textsc{sun}]\!]$: the antecedent remains true throughout

Unification: Both flashbacks/explanations and counterfactuals involve:

1. Editing an established timeline (via $\textsc{insert}$ or $\textsc{replace}$)
2. Seeking coherence via the same ODE dynamics
3. Respecting linguistic constraints via the Anchor term

The difference: $\textsc{insert}$ adds factual content; $\textsc{replace}$ substitutes hypothetical content

Energy (=Incoherence / Tension) Based Temporal Resolution

When linguistic cues underdetermine temporal relation, we disambiguate by choosing among candidate trajectories.

Example: "Bob won. He practiced." (bare past compatible with narration or explanation)

- Narration: $\langle [\![\text{WIN}]\!], [\![\text{PRACTICE}]\!] \rangle$ (practiced after winning?)
- Explanation: $\langle [\![\text{PRACTICE}]\!], [\![\text{WIN}]\!] \rangle$ (practiced before winning)

Energy (=Incoherence / Tension) based resolution: For each candidate $\bar{X}_c$, compute stabilized interpretation and its energy. Prefer lowest energy / most coherent candidate:

$$\boxed{\bar{X}^\star = \arg\min_{\bar{X}_c} H\left(\mathcal{I}^{W,\theta}(\bar{X}_c)\right)}$$

If W encodes "practice typically precedes winning," the explanation trajectory is more coherent / has lower energy $\Rightarrow$ preferred interpretation

The **Gibbs distribution** $P(\bar{X}) \propto \exp(-H(\bar{X}))$ assigns probabilities to candidates: high-energy (incoherent) $\rightarrow$ low probability; low-energy (coherent) $\rightarrow$ high probability

Window of Plasticity and Defeasibility

Strict incrementality: Once stabilized, $X_t^\star$ becomes the new ceiling—inferences "harden"

Alternative: Window of plasticity for recent positions, retain both:

- X_t^{initial}: what was *said* (original linguistic constraints)
- $X_t^\star$: what is currently *understood* (after inferential enrichment)

Explains defeasibility (Asher, 2013):

John fell. Mary pushed him. He rolled off the cliff.

First two sentences invite Explanation (push caused fall). Third sentence shifts to Narration (fall, push, roll). Retaining X^{initial} allows revision.

Implications for counterfactuals:

- Compare CF antecedent against X^{initial} (what was said)?
- Or against $X^\star$ (what is understood, including inferences)?
- Different predictions for how freely CF scenarios can diverge from actuality