Patchi: Formalizing and Stress-Testing Pygmalion's Relation-Based Theory of Cognition

Theory: Pygmalion. **Formalization, implementation, and experiments:** Emma Leonhart with an autonomous coding agent.

Abstract

Pygmalion's notebook *artificial time* sketches an ambitious unified theory of machine cognition: meaning is made of relations between words; words are agreements on labels; context is the base relation from which all others draw meaning; information is carried by *infons* in *situations*; memory is a recursion indexed by an "artificial time"; words map bijectively to reconfigurable "VHDL-style" neural blocks; and a topos-level layer reasons about those blocks. We do two things with it. First, we read the sprawling notebook as a single *layered stack* and ground each layer in established theory (situation semantics, distributional and compositional vector semantics, vector-symbolic architectures and conceptual spaces, neuro-symbolic AI and knowledge-graph embeddings, and topos/modal/spatial/temporal logic), finding that the layers are well-trodden and the *bridges between them* are where the originality and the risk live. Second, we build a runnable reference implementation of the core and test its most distinctive component — a similarity-weighted blending operator — on both a controlled synthetic task and real GloVe embeddings against the human WordSim-353 judgements. The result is sharply regime-dependent and, on real data, negative: blending *denoises* noisy vectors (up to +0.14 Spearman over the raw vectors) but *hurts* clean pretrained embeddings (raw GloVe scores 0.50, every reconstruction variant 0.43–0.45). The operator is a noise-conditional smoother, not a free improvement over standard embeddings. We report the negative result in full, because it is the opposite of what the synthetic run alone would have implied.

1. Introduction

This paper formalizes and empirically stress-tests a theory of cognition due to **Pygmalion**. The source material — a research notebook and an accompanying report — proposes that cognition can be assembled bottom-up from relations between words, with a stack of increasingly abstract structures built on top. The notebook is idiosyncratic and wide-ranging; read charitably it is not a grab-bag but a coherent architecture that recruits one established intellectual tradition per layer.

Our contribution is twofold: (i) a literature-grounded reading of the framework that separates what is already known from what is genuinely new, and (ii) a working implementation of the core plus an honest empirical test of its signature operator, including a negative result on real data.

2. The framework as a layered stack

The notebook's own closing hierarchy makes the stack explicit:

```
infons / infon-classes        → situation theory (Devlin)
WordClasses (poly+geom+axiom)  → conceptual spaces + VSA
```

```
(Gärdenfors, Plate/Kanerva)
neural "VHDL" blocks           → distributional / compositional
semantics (word2vec, DisCoCat)
spatial / temporal logic       → modal / spatial / temporal logic
(Kripke, Aiello, Pnueli)
objects → thought → idea        → emergent (the project's own
contribution)
collectives / topos            → categorical logic (Mac Lane–
Moerdijk, Goldblatt)
```

Each arrow is a *bridge* Pygmalion asserts: “spatial logic is the gateway to vector algebra”; “topos as VHDL category reasoning”; the bijective “translator: WordClass ↔ NeuralCircuitBlock”. The layers are mature prior art; the bridges are the research, and several are not known to be consistent.

3. Related work (what is already known)

- **Meaning from distribution.** A word’s meaning is fixed by the company it keeps (Firth 1957; Harris 1954) and is representable geometrically, with relations as consistent vector offsets (Mikolov et al. 2013; Pennington et al. 2014; Bordes et al. 2013). Pygmalion’s “meaning as relations between words” and the “king − man + woman ≈ queen” example are this tradition.
- **Information as infons-in-situations.** The `⟨relation, args, polarity⟩` infon and the support relation `s ⊨ σ` are Devlin’s situation theory (Barwise & Perry 1983; Devlin 1991) verbatim; the symbolic account is complete, a graded/learned one is not.
- **Compositional vectors, categorically.** DisCoCat (Coecke, Sadrzadeh & Clark
2010. already fuses grammar, category theory, and distributional vectors to compose sentence meaning — the closest existing realization of the vision.
- **Binding and regions.** VSA/HDC (Plate 1995; Kanerva 2009; Gayler 2003; Smolensky 1990) supplies binding/bundling; Gärdenfors’ conceptual spaces (2000) supply concepts-as-regions — Pygmalion’s “classes as perimeters.”
- **Worlds, space, time, topoi.** Kripke (1963), the *Handbook of Spatial Logics* (Aiello et al. 2007), Pnueli (1977), and topos internal logic (Mac Lane & Moerdijk 1992; Goldblatt 1979) each exist in mature form.

The recurring shape of the gap: the components are known in isolation; the bridges are not standard. The full literature review with ~26 cited sources is in the repository (`literature/REVIEW.md`).

4. The implemented system

We built a runnable Python core (`patchi`) covering a vertical slice of the stack, each component unit-tested:

- **WordClass lexicon** — words as vectors with a parameter bag and cosine nearest-neighbour lookup.
- **Signed relation graph** — synonym(+)/antonym(−) edges (the stimulator/ inhibitor spectrum) with TransE-style relation offsets; held-out edges are recovered by offset arithmetic.
- **Similarity-weighted blending operator** — `blend(w) = Σ sim(w,sᵢ)^p·vec(sᵢ) / Σ sim(w,sᵢ)^p`, with uniform weighting recovering an additive baseline.
- **Infon/situation layer** — `⟨relation, args, polarity⟩` with a *graded* `support(s,σ) ∈ [0,1]` (Devlin’s binary support made continuous and vector-backed), so context-conditioning measurably changes outputs.
- **Proof(walk) trace** — every output records the words/weights that

produced it, with a single-source-of-truth discipline so the trace cannot diverge from the computed value.
- **Reduced cores for the harder bridges** — a registry-backed bijective translator (bijection over the lexicon, grown on demand), a category of invertible affine blocks with property checkers (the computable shadow of the topos layer), and "artificial time" as the discrete recursion index of a memory cell. These are honest reductions; the full versions are named as future work.

5. Methods

We test the blending operator — the framework's most distinctive composition primitive — against two baselines (raw vectors; additive = unweighted neighbour mean) on two benchmarks, using the *same* operator code in both.

- **Synthetic denoising.** Clustered prototype vectors plus Gaussian noise; ground truth = cosine of the clean prototypes. Score = Spearman(method pairwise cosine, ground-truth cosine).
- **Real embeddings.** GloVe-50 (top 100k words) vs the human **WordSim-353** similarity judgements (all 353 pairs). Score = Spearman(method pair similarity, human score).

6. Results

Real embeddings (the headline). Raw GloVe-50 scores Spearman **0.5033** — matching the known literature value, a check that the harness is correct. Every reconstruction underperforms it:

k	power	additive	blend	blend − raw
3	6	0.4420	0.4470	−0.0564
5	2	0.4309	0.4336	−0.0697
10	2	0.4318	0.4347	−0.0686
25	2	0.4228	0.4256	−0.0777

On clean pretrained vectors, reconstructing a word from its neighbourhood **loses to doing nothing** (best blend 0.447 vs raw 0.503). The similarity weighting reliably beats the unweighted average, but only by +0.001 to +0.009.

Synthetic (the contrast). When vectors are noisy, reconstruction *denoises*:

noise	power	raw	additive	blend
0.4	4.0	0.892	0.829	**0.972**
0.8	4.0	0.689	0.751	0.866
1.6	4.0	0.363	0.352	0.349

Here blend beats raw by up to +0.14, and the weighting beats additive by +0.10–0.14 — until noise is high enough that the neighbourhood itself is unreliable and the gain vanishes or goes slightly negative.

7. Discussion

The blending operator's value is **entirely conditional on input noise**. It is a denoiser: it wins exactly when averaging over trustworthy neighbours recovers signal (synthetic, low–moderate noise) and loses when the base vectors are already clean (GloVe), where averaging washes out discriminative information. The similarity

weighting — the operator's distinctive ingredient over a flat average — is real but small in both regimes, never the difference between winning and losing; the decision *to reconstruct at all* dominates. This is the opposite of what the synthetic run alone would have suggested, which is precisely why running the real benchmark mattered.

8. Limitations

One embedding model (GloVe-50) and one dataset (WordSim-353); SimLex-999, larger dimensions, and word2vec/fastText would qualify the result. Reconstruction here replaces a word entirely by its neighbourhood — a residual form (`raw + α·blend`) is untested. The harder bridges (full topos internal logic, richer polynomial/geometric block internals, the control-system reframing of neural nets) are implemented only as reduced cores or named as future work, not papered over.

9. Conclusion

Pygmalion's framework is a coherent stack whose layers are well-grounded and whose bridges are the open contribution. Its signature blending operator, when actually measured, is a noise-conditional smoother rather than a free improvement over standard embeddings — a narrow but real finding, reported here including the negative result on clean data. The full implementation, literature review, and reproducible benchmarks are public at https://github.com/EmmaLeonhart/patchi (report: https://emmaleonhart.github.io/patchi/).

References

Selected; full notes with verified identifiers in `literature/sources.md`. Barwise & Perry, *Situations and Attitudes* (1983). Devlin, *Logic and Information* (1991). Firth (1957); Harris (1954). Turney & Pantel (2010). Mikolov et al. (2013); Pennington et al. (2014). Coecke, Sadrzadeh & Clark (2010). Plate (1995); Kanerva (2009); Gayler (2003); Smolensky (1990). Gärdenfors (2000). Bordes et al. (2013, TransE). Vaswani et al. (2017). Mac Lane & Moerdijk (1992); Goldblatt (1979). Kripke (1963). Aiello, Pratt-Hartmann & van Benthem (2007). Pnueli (1977).