Rational Communication Shapes Morphological Composition

Fengyuan Yang, Yongqian Peng, Yuxi Ma, Chenheng Xu, and Yixin Zhu ✉
Peking University

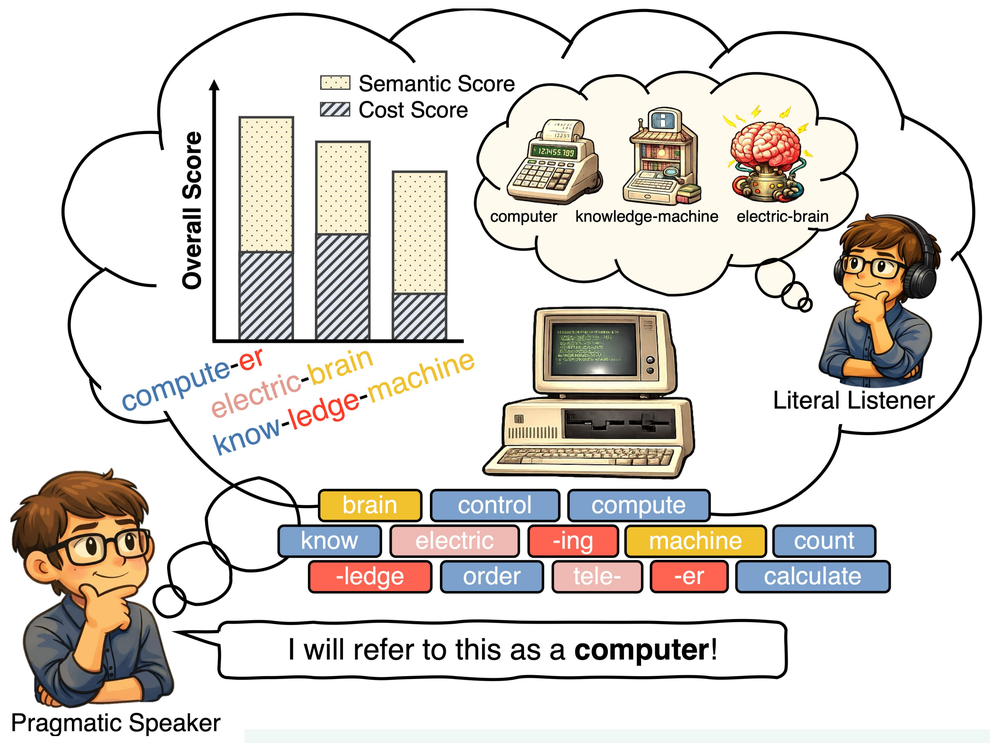
TLDR: We model word formation as a pragmatic speaker choice: selecting morphemes that maximize listener recoverability while minimizing production cost. Across 4,323 English compounds and derivations (1820–2019), an RSA-style model predicts attested forms over contemporaneous alternatives.

Why do languages pick this morpheme combination?

Languages constantly invent new words by combining existing parts.

computer (comput-er) English	**电脑** (electric-brain) Chinese	**tietokone** (knowledge-machine) Finnish

All three combinations were morphologically available in all three languages. Why do languages prefer one combination over others?



Composition as a pragmatic speaker's choice

Given a target concept c and time t, which combination among candidates $C(c,t)$ is communicatively optimal or "rational"?

"Listener" (semantic compatibility)
- Needs to infer meaning from morphemes
- Prefers clarity and recoverability

Qwen3-8B embeddings morph–concept similarity

$$L_0(c \mid u, \mathcal{L}_t) \propto S_0(u \mid c, \mathcal{L}_t) P(c \mid \mathcal{L}_t)$$

"Speaker" (production cost)
- Prefers low effort, short forms
- Prefers common, familiar morphemes

word length · frequency · phoneme and syllable count

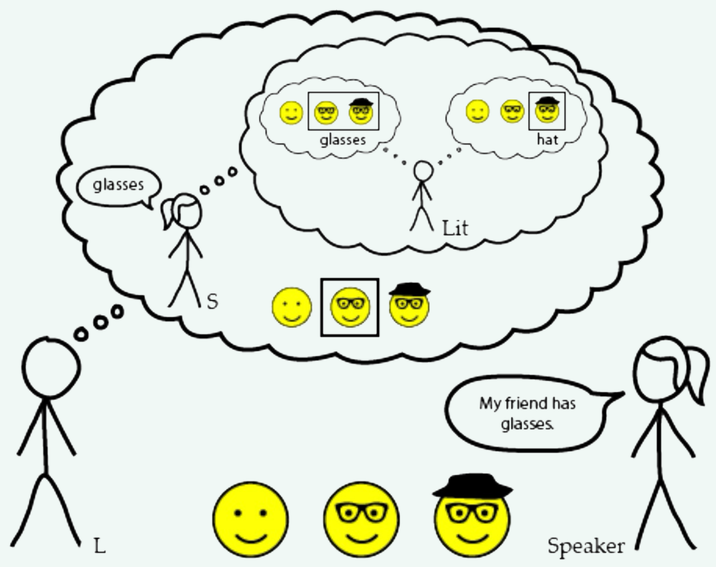
$$\mathrm{Cost}(u, \mathcal{L}_t) = \sum_{i=1}^{m} h_\phi\big(\mathrm{feat}(\mu_i, \mathcal{L}_t)\big)$$

"Pragmatic speaker S_1" (Integrating informativeness and cost)

$$S_1(u \mid c, \mathcal{L}_t) \propto \exp\big(\log L_0(c \mid u, \mathcal{L}_t) - \mathrm{Cost}(u, \mathcal{L}_t)\big).$$

Rational speech act framework

(Frank & Goodman, 2012; Goodman & Frank, 2016)



Material and methods

Time-indexed lexicon $\mathcal{L}_t$

separate word2vec per COHA decade · COCA year; candidates from synset neighbors + kNN

Dataset: 4,323 lexicalized compositions; 1820–2019.

Metrics: Mean Reciprocal Rank (MRR), top-k accuracy (Acc@k)

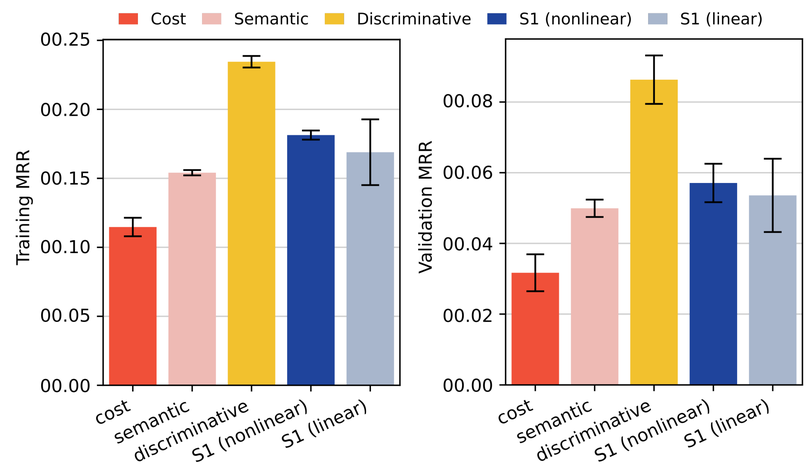
Cost	*Cost only*
Semantic	*S_0 only*
Discriminative	*full 13-dim MLP*
Nonlinear S_1	*MLP over frozen scores*
Linear S_1	*mirrors RSA utility*

```
Learned weights
   α (semantic) = 0.38
   β (cost)     = 0.48
   β/α ≈ 1.28
```

Results

S_1 beats both single-factor baselines on MRR and $Acc@k$, → informativeness and cost each contribute independently.

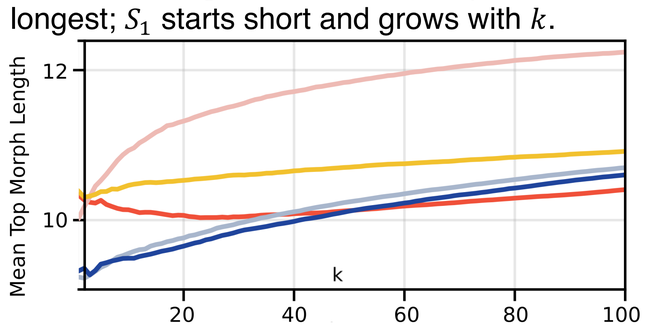
Semantic recoverability dominates (MRR 0.047), but adding cost yields a consistent further gain (MRR 0.050–0.053).



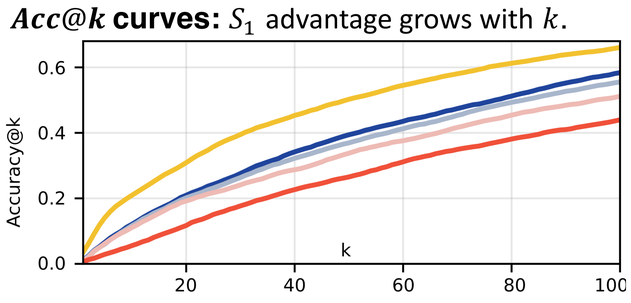
Semantic-only over generates **redundant** forms *(e.g. "cynic-cynic" for cynicism)*; Cost-only picks **cheap but meaningless** fragments *(e.g. "ed-ing" for saucepan)*; S_1 avoids both failure modes.

Target	Model	Top-3 predictions	Rank
fiancee	Linear S_1	fiance-eld, fiance-ee, fianceive	#2
	Semantic	fiance-est, fiance-eld, fiance-person	#5
	Cost	eld-ly-ity, ed-ency-ly, from-er-ly	#74

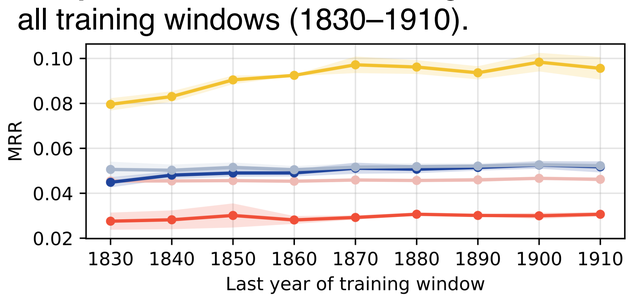
Predicted length vs k: Semantic predictions longest; S_1 starts short and grows with k.



$Acc@k$ curves: S_1 advantage grows with k.



Temporal robustness: Ordering stable across all training windows (1830–1910).



Limitations

- Current model ignores morpheme ordering, morphotactic constraints, and head-modifier structure;
- Semantic representations are contemporary rather than historical;
- Evaluation is restricted to English.

Future directions

- Morphotactic composition models;
- Diachronic semantic representations;
- Typologically diverse languages;
- Behavioral experiments on novel word formation.