Rational Communication Shapes Morphological Composition

Fengyuan Yang[1,2,3,4,5], **Yongqian Peng**[1,2,3,4,5], **Yuxi Ma**[1,2,4,5], **Chenheng Xu**[1,2,4,5], **and Yixin Zhu**[2,1,4,5]✉

[1] Institute for Artificial Intelligence, Peking University [2] School of Psychological and Cognitive Sciences, Peking University
[3] Yuanpei College, Peking University [4] State Key Laboratory of General Artificial Intelligence, Peking University
[5] Beijing Key Laboratory of Behavior and Mental Health, Peking University

Abstract

Human languages expand vocabularies by combining existing morphemes rather than inventing arbitrary forms. Communicative efficiency shapes lexical systems at multiple levels (Gibson et al., 2019), yet morphological composition—combining morphemes through compounding or affixation—has rarely been modeled as a historically situated speaker choice among competing morpheme sequences, leaving unanswered why a language settles on one morpheme combination over other plausible alternatives. We ask whether a trade-off between listener recoverability and speaker production cost can predict attested compositions over contemporaneously available alternatives. Here we show, within the Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016) using a time-indexed lexicon constructed from Corpus of Historical American English (COHA) and Corpus of Contemporary American English (COCA), that across 4323 naturally occurring English compounds and derivations spanning 1820–2019, attested compositions are systematically ranked above unattested alternatives generated from contemporaneously available morphemes. Models integrating semantic informativeness with production cost outperform semantic-only and cost-only baselines on Mean Reciprocal Rank (MRR) and top-k accuracy (Acc@k), with the advantage of Pragmatic Speaker model (S_1) over the semantic-only baseline growing as the candidate set expands, where meaning alone leaves morphological choice underdetermined. These findings suggest that lexicalization reflects a communicative trade-off between expressiveness and efficiency, extending rational accounts of communication from utterance-level choice to the internal structure of words.

Keywords: morphology; composition; rational communication; computational modeling

Introduction

Human languages are shaped by the pressure to communicate efficiently: across syntax, semantics, and the lexicon, language structure reflects a systematic tendency to convey meaning while minimizing effort (Gibson et al., 2019; Jiang et al., 2024, 2025; Salge et al., 2015; Zipf, 1949). This pressure is visible at the lexical level, where word lengths correlate not with usage frequency alone but more precisely with contextual predictability (Piantadosi et al., 2011): words that are easier to anticipate tend to be shorter, as an efficient coding system would predict (Jaeger, 2010; Levy, 2008; Shannon, 1948). Word formation brings this pressure into sharp focus: when speakers need to name new concepts, they draw on existing morphemes, combining them via compounding or affixation to produce words whose meanings are recoverable from their parts (Algeo, 1980; Brinton & Traugott, 2005; Xu et al., 2024).

We use *morphological composition* to refer specifically to the combinatorial assembly of morphemes into new words, including compounds and derivations. This is a central subtype of *word formation*, a broader domain that also includes backformation, blending, and conversion (Bauer, 2001; Peng et al., 2025; Plag, 1999; Štekauer & Lieber, 2005). Morphological composition is highly productive: speakers can coin new words by combining morphemes, and listeners can often interpret the results without prior exposure when component meanings are transparent. Experimental and computational studies confirm that semantic transparency systematically affects how novel compounds and derivations are interpreted (Levin et al., 2019; Marelli & Baroni, 2015; Mattiello & Dressler, 2018; Reddy et al., 2011). Yet strikingly different morpheme combinations across languages can name the same concept: English *computer* uses *comput-er*, while Chinese translates directly as *electric-brain* and Finnish *tietokone* as *knowledge-machine* (see Fig. 1). These divergences reflect not only what morphemes are available in a language's inventory but also what a listener can plausibly infer from them. Morphological composition is therefore not merely a structural property of language; it is a communicative choice made under efficiency pressures.

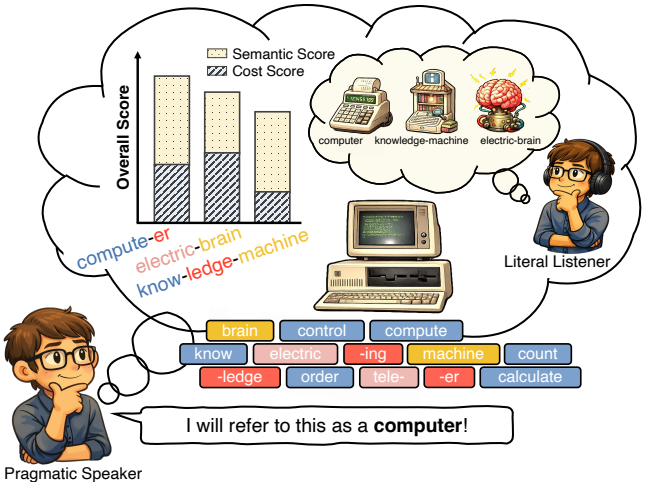

Figure 1: **Modeling morphological composition as rational communication.** A pragmatic speaker (left) constructs a word by selecting among candidate morpheme combinations available in the lexicon, balancing semantic informativeness—how well a literal listener (right) can infer the intended meaning from competing alternatives—against production cost. The example shows alternative compositions for the concept "a programmable machine that performs arithmetic or logical operations," including English *compute-er*, Chinese *electric-brain*, and Finnish *knowledge-machine*.

Prior computational and quantitative work already shows that morpheme selection is systematic rather than arbitrary. Morphological productivity research links the availability of word-formation patterns to lexical statistics, semantic trans-

parency, and phonological or structural constraints (Arndt-Lappe, 2015; Bauer, 2001; Plag, 1999). Analogical models capture competition among suffixation patterns such as *-ity vs. -ness* through gradient generalization over lexical exemplars (Arndt-Lappe, 2015). Information-theoretic frameworks characterize global lexical systems as efficient encodings of semantic structure (Gibson et al., 2019; Zaslavsky et al., 2018). Closest to our work, Xu et al. (2024) show that word reuse and combination support efficient communication of emerging concepts, but their model evaluates corpus-level lexical adoption rather than the speaker-side choice among alternative morpheme sequences for the same target meaning.

These frameworks illuminate important aspects of word formation, but leave open the candidate-level question we pursue here. Productivity and analogy models explain which word-formation patterns are available or preferred, not which particular morpheme sequence should be chosen when several could name the same concept. Information-theoretic accounts evaluate the efficiency of a lexical code as a whole rather than ranking specific morphological candidates available at a particular historical moment. Xu et al. (2024) ask which introduced forms survive in the lexicon, whereas we ask which morpheme combination is communicatively optimal as a historically situated choice. Addressing this requires a framework with an explicit candidate-comparison structure: one that ranks alternative morpheme sequences by how recoverable each is for a listener and how costly each is for a speaker.

The RSA framework (Frank & Goodman, 2012; Goodman & Frank, 2016) provides exactly this. A S_1 selects among explicit form alternatives by reasoning about listener recoverability and production cost, producing a ranked distribution over candidates via a utility U that trades informativeness against cost, while a Literal Speaker model (S_0) supplies the semantic compatibility baseline from which pragmatic reasoning departs. This does not mean that RSA should replace global efficiency, analogy, or adoption models; rather, it offers a complementary and interpretable decomposition for the moment-specific choice among candidate forms. RSA has been applied to referential communication (Goodman & Frank, 2016), pro-drop (Chen et al., 2018), and cooperative explanation (Chandra et al., 2024), yet morphology has remained largely untouched.

In this paper, we extend RSA-style pragmatic modeling to morphological composition, treating word formation as a cooperative referential game. For a target concept c at time t, a S_1 selects among candidate morpheme sequences $u \in C(c,t)$ by assigning utility $U(u,c)$ that trades listener recoverability against speaker production cost. Candidates are represented by per-morpheme-to-concept similarity statistics rather than a single composite embedding, preserving compositional structure while remaining order-agnostic. Using a time-indexed lexicon $\mathcal{L}_t$ built from WordNet (Miller, 1995) and historical COHA/COCA statistics, we evaluate 4323 English compounds and derivations spanning 1820–2019. Models that combine informativeness and cost outperform single-factor baselines, with the advantage growing as the candidate space expands.

Computational Framework

We treat existing poly-morphemic words as traces of historical lexical choices made around emergence time t. Empirically, t is the first year with a non-zero COHA/COCA count; in the model, it approximates the lexical state near conventionalization. A time-indexed lexicon $\mathcal{L}_t$ supplies the shared linguistic background: word meanings, historical frequency, phonology, and form features. The central question is whether attested morpheme sequences rank above contemporaneous alternatives under a trade-off between recoverability and cost.

Notation. Let c be a target concept, $u = (\mu_1, \ldots, \mu_m)$ a candidate morpheme sequence, and $C(c,t)$ the candidate set available at time t. S_0 denotes semantic compatibility, Literal Listener model (L_0) listener recoverability, S_1 pragmatic speaker preference, and $\text{Cost}(u, \mathcal{L}_t)$ production cost. These are population-level operationalizations of RSA terms, estimated from historical corpus resources rather than individual judgments. For a concept at lexical emergence, S_1 conveys c using morphemes from $\mathcal{L}_t$, balancing listener recoverability against production cost. We formalize listener interpretation, cost, and speaker choice in turn.

Inferring listener interpretation

To infer $L_0(c \mid u, \mathcal{L}_t)$, we define a literal production model $S_0(u \mid c, \mathcal{L}_t)$ capturing compatibility between utterance and concept. Following Goodman and Frank (2016), listener interpretation is:

$$L_0(c \mid u, \mathcal{L}_t) \propto S_0(u \mid c, \mathcal{L}_t)\, P(c \mid \mathcal{L}_t), \tag{1}$$

where $P(c \mid \mathcal{L}_t)$ is a concept prior. Since our targets are novel concepts without dedicated lexical entries, we assume a uniform prior over candidate concepts.

Rather than explicitly normalizing this posterior, we learn a compatibility function $f_\theta(u,c)$ intended to be monotonic with listener recoverability. Concretely, we compute similarities between the target concept embedding and the embeddings of morphemes in u. Rather than representing a candidate with a single composite embedding, we summarize the distribution of similarities between each morpheme and the concept using five statistics: mean, maximum, standard deviation, entropy, and the similarity between the whole candidate and the target. These feed the learned semantic scorer f_θ.

To keep the search space tractable while preserving semantic plausibility, a gate $g(u,c)$ restricts candidate morphemes to a neighborhood of the target gloss d_c. Let $\text{kNN}(d_c)$ denote the morphemes among the k nearest semantic neighbors of d_c:

$$g(u,c) = \begin{cases} 0, & \text{if } \exists\, \mu \in u \text{ such that } \mu \notin \text{kNN}(d_c), \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

Candidates with $g(u,c) = 0$ are excluded before ranking, giving the gated compatibility function:

$$\tilde{f}_\theta(u,c) = \begin{cases} f_\theta(u,c), & \text{if } g(u,c) = 1, \\ -\infty, & \text{if } g(u,c) = 0. \end{cases} \tag{3}$$

Estimating production cost

Production cost is estimated from corpus- and form-based features in $\mathcal{L}_t$, including morpheme frequency, phonological complexity, and length. A learned function h_ϕ maps each morpheme's feature vector to a scalar cost, and candidate cost is additive:

$$\text{Cost}(u, \mathcal{L}_t) = \sum_{i=1}^{m} h_\phi\big(\text{feat}(\mu_i, \mathcal{L}_t)\big), \tag{4}$$

where $\text{feat}(\mu_i, \mathcal{L}_t)$ extracts the time-indexed feature vector for morpheme μ_i. This treats frequent, shorter, and phonologically simpler morphemes as less costly, consistent with established links between predictability, form complexity, and production effort (Jaeger, 2010; Levy, 2008).

Integrating informativeness and cost

Unlike S_0, S_1 selects utterances by explicitly trading listener recoverability against production cost. In standard RSA form:

$$S_1(u \mid c, \mathcal{L}_t) \propto \exp\big(\log L_0(c \mid u, \mathcal{L}_t) - \text{Cost}(u, \mathcal{L}_t)\big). \tag{5}$$

We implement this trade-off via a utility $U_\theta(u, c)$ defined over two learned base scores: gated semantic compatibility $\tilde{f}_\theta(u, c)$, approximating listener informativeness, and production cost $\text{Cost}(u, \mathcal{L}_t)$. The linear model combines these as:

$$U_\theta(u, c) = \alpha \cdot \tilde{f}_\theta(u, c) - \beta \cdot \text{Cost}(u, \mathcal{L}_t) + b, \tag{6}$$

where α, β, and b are learnable. The nonlinear model uses the same two inputs but allows interactions between informativeness and cost via a small MultiLayer Perceptron (MLP):

$$U_\theta(u, c) = \text{MLP}_\theta\big(\tilde{f}_\theta(u, c),\ \text{Cost}(u, \mathcal{L}_t)\big). \tag{7}$$

We refer to these as the Linear S_1 and Nonlinear S_1 models. The linear version closely follows the standard RSA utility; the nonlinear version permits richer interactions between the two factors while retaining the same two-factor decomposition.

Taken together, the framework casts morphological composition as candidate ranking under a communicative trade-off: attested forms should rank highly when evaluated for recoverability and efficiency relative to alternatives in $\mathcal{L}_t$.

Materials and Methods

Dataset

The dataset contains 4323 English compounds and derivations from WordNet (Miller, 1995), spanning first appearances from 1820 to 2019. Each item has a gloss d_c, a gold morpheme sequence segmented with MorphSeg (TheWelcomer, 2025), and an emergence time t defined as the first year with non-zero COHA/COCA count. First attestation is an imperfect proxy because lexicalization may lag concept emergence (Brinton & Traugott, 2005); phoneme and syllable counts come from the CMU Pronouncing Dictionary.[1]

The morpheme lexicon combines MorphSeg morphemes with an affix inventory (Affixes.org, 2008), assigning definitions from WordNet or the affix dictionary. Morpheme availability is quantified using cumulative type and token frequencies by decade for COHA (Davies, 2012) and by year for COCA (Davies, 2010).

Candidate sets $C(c, t)$ combine morphemes from WordNet synset lemmas, relational neighbors, and nearest neighbors from time-indexed word2vec spaces (Mikolov et al., 2013). We train skip-gram word2vec models separately on COHA decade slices (1820–2010) and COCA yearly slices (1990–2019), so that each target year draws distributional neighbors from its corresponding historical embedding space. We enumerate up to three morphemes per candidate, capped by length and candidate count, and apply the semantic gate $g(u, c)$ to avoid exhaustive enumeration over the full inventory.

Feature representations

Semantic features. Target glosses and morphemes are embedded with Qwen3-Embedding-8B (Zhang et al., 2025), using both surface forms and available definitions for morpheme representations. For each candidate, the distribution of morpheme–concept cosine similarities is summarized by five statistics: mean, maximum, standard deviation, entropy, and the similarity between the whole candidate and the target. These form the 5-dimensional input to f_θ.

Cost features. Each morpheme receives eight time-indexed features: character length, cumulative type and token frequency, 30-year token frequency, standalone and 30-year standalone word frequency, phoneme count, and syllable count. Frequency-derived features are clipped at zero and log-transformed before model input. All features are indexed by the target word's emergence year t.

Model designs

We evaluate five ranking models grouped into four families: cost-only, semantic-only, discriminative, and two S_1 variants. All models are trained to rank the gold morpheme sequence above alternatives for the same target concept, with batches padded and masked so that only valid candidates and morpheme positions contribute.

Cost model. A two-layer per-morpheme MLP h_ϕ maps each time-indexed cost feature vector $\mathbf{x}(\mu_i, t)$ to a scalar production cost. The candidate score is the negated sum over morphemes, so higher scores correspond to lower estimated cost:

$$\text{score}_{\text{cost}}(u) = -\gamma \sum_{i=1}^{m} h_\phi\big(\mathbf{x}(\mu_i, t)\big),$$

where $\gamma > 0$ is a learnable scale parameter.

Semantic model. A three-layer MLP f_θ maps the normalized semantic statistic vector $\mathbf{s}(u, c)$ to a scalar compatibility score, approximating S_0 and serving as the single-factor informativeness baseline:

$$\text{score}_{\text{sem}}(u, c) = f_\theta\big(\text{LN}(\mathbf{s}(u, c))\big)/\tau,$$

where $\tau > 0$ is a learned temperature that calibrates ranking confidence.

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict; version 0.7b.

Discriminative model. Semantic statistics and masked-mean-aggregated cost features (5- and 8-dimensional, respectively) are concatenated into a 13-dimensional vector and scored by a single three-layer MLP. This model serves as a flexible upper bound: it accesses all available information simultaneously but does not impose an interpretable informativeness–cost decomposition.

S_1 models. Rather than consuming raw features, the S_1 models operate as second-stage models over independently trained, *frozen* Cost and Semantic base scorers. Freezing keeps the two RSA components identifiable and prevents the S_1 model from collapsing into an unconstrained discriminative ranker. The frozen scalar outputs are batch-normalized to a common scale and combined by a learned function ψ:

$$\text{score}_{S_1}(u, c) = \psi\big(\text{score}_{\text{cost}}(u),\ \text{score}_{\text{sem}}(u, c)\big).$$

A *linear* variant directly mirrors the RSA utility $U = \alpha \log L_0 - \beta\,\text{Cost}$, while a *nonlinear* variant uses a small two-layer MLP to capture interaction effects between the two factors.

Training and evaluation

All models are trained as rankers with a pairwise softplus loss that encourages the gold morpheme sequence to score above sampled negatives, selected via curriculum hard-negative mining that emphasizes semantically similar candidates and those overlapping with the gold sequence. Models are optimized with Adam, gradient clipping, and early stopping on validation performance.

Data are split by year-stratified holdout to ensure historical periods remain represented in each partition: 20% of items within each year are held out for testing, and the remainder are divided into training and validation using one fold of a 5-fold rotation scheme, yielding 864 held-out test items. Training and validation MRR are used for fitting diagnostics and model selection. Final test metrics are computed on the held-out partition by ranking the gold sequence among up to 1024 candidates sampled from $C(c, t)$, reporting MRR and Acc@k.

Results

We first report held-out ranking performance, then use training and validation curves as diagnostics. Higher MRR and Acc@k indicate that the attested morpheme sequence is assigned a better rank.

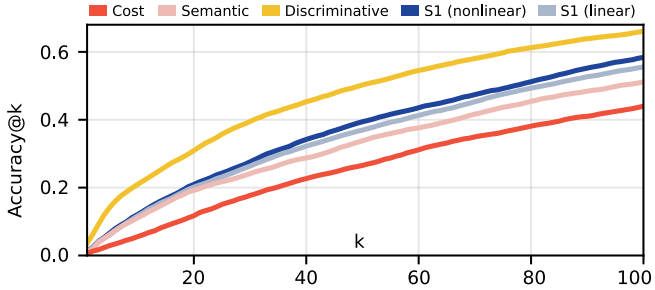

Figure 2: **Held-out Acc@k as a function of k.** The discriminative model consistently identifies the gold morpheme sequence most accurately. The advantage of S_1 over the semantic-only baseline grows with k, consistent with pragmatic reasoning becoming most useful when many semantically plausible alternatives compete.

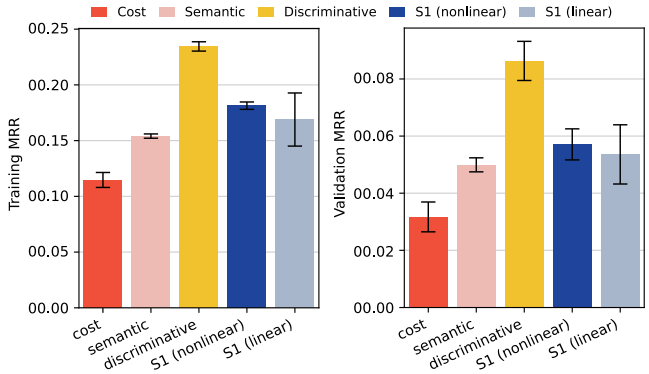

Figure 3: **Training and validation MRR diagnostic.** The same broad ordering appears during fitting: models with access to semantic information outperform the cost-only baseline, and models that combine both sources of information perform best among the interpretable variants. Error bars indicate standard deviation across repeated training runs.

Quantitative results

Fig. 2 summarizes held-out ranking performance across retrieval thresholds. The discriminative model remains strongest overall, but the theoretically important comparison is between the S_1 models and the single-factor baselines. The S_1 curves generally sit above both semantic-only and cost-only, with the clearest separation from semantic-only emerging at larger k: the regime in which many candidates are meaning-adjacent, semantic salience alone leaves the choice underdetermined, and production cost helps select among plausible forms.

The scalar MRR results follow the same ordering. Cost alone is weakest (MRR = 0.031), semantic information is stronger (MRR = 0.047), and adding cost yields a modest further gain for the S_1 models (MRR = 0.050–0.053). This gain is also visible at concrete retrieval thresholds: the nonlinear S_1 model reaches Acc@10 = 12.08% and Acc@20 = 20.94%, compared with 11.10% and 19.19% for the semantic-only baseline. The discriminative model performs best overall (MRR = 0.096), indicating that additional feature interactions remain beyond the constrained two-scalar S_1 decomposition.

Training and validation behavior provides a diagnostic check on this held-out pattern. As shown in Fig. 3, the same broad ordering appears during fitting: semantic information improves over cost alone, S_1 improves over the single-factor baselines among the interpretable models, and the discriminative model remains strongest.

Qualitative results

Tab. 1 illustrates the complementary failure modes of the single-factor baselines. The cost-only model favors short, high-frequency fragments—for instance, `ed+ing` for *saucepan*—that are cheap but semantically uninformative. The semantic-only model errs in the opposite direction: for *cynicism* it prefers the redundant `cynic+cynic`, and for *fiancée* it selects

Table 1: **Qualitative comparison of model rankings.** Gold morpheme sequences alongside the top-3 candidates produced by three models. The S_1 model ranks the gold sequence at or near the top; the cost-only model selects high-frequency but semantically irrelevant fragments; the semantic-only model selects meaning-adjacent but morphologically odd forms.

Model	Gold	Top-3 predictions	Rank
Linear S_1	[‘laundr’, ‘y’]	[‘laundr’, ‘y’], [‘laundr’, ‘ify’, ‘ing’], [‘ca’, ‘laundr’, ‘launder’]	1
Linear S_1	[‘affiliate’, ‘ed’]	[‘affiliate’, ‘ation’], [‘affiliate’, ‘ed’], [‘affiliate’, ‘ation’, ‘ed’]	2
Linear S_1	[‘fiance’, ‘ee’]	[‘fiance’, ‘eld’], [‘fiance’, ‘ee’], [‘fiance’, ‘ive’]	2
Semantic	[‘fiance’, ‘ee’]	[‘fiance’, ‘est’], [‘fiance’, ‘eld’], [‘person’, ‘fiance’]	5
Semantic	[‘cynic’, ‘ism’]	[‘cynic’, ‘cynic’], [‘cynic’, ‘antipathy’], [‘cynic’, ‘ance’]	83
Cost	[‘sauce’, ‘pan’]	[‘ed’, ‘ing’], [‘ed’, ‘ing’, ‘ing’], [‘y’, ‘ing’, ‘ation’]	74
Cost	[‘fiance’, ‘ee’]	[‘eld’, ‘ly’, ‘ity’], [‘ed’, ‘ly’, ‘ency’], [‘ly’, ‘from’, ‘er’]	81

semantically related but morphologically odd forms such as `person+fiance`.

The S_1 model better balances these pressures, ranking *laundry* first and placing the gold forms for *affiliated* and *fiancée* at rank 2, with near-misses that differ mainly in suffix choice. These examples show how integrating recoverability and cost favors candidates that are meaning-aligned without becoming unnecessarily long or morphologically implausible.

Discussion

The results support a conservative version of the rational-communication hypothesis. Semantic recoverability is the strongest single cue, but attested morphological forms rank better when recoverability is evaluated together with production cost. The effect is modest, as expected for a historically noisy word-formation problem with large candidate sets, but it is consistent across ranking metrics, qualitative examples, and temporal robustness checks. The discriminative model's advantage shows that lexicalization depends on additional structure beyond what the current RSA operationalization captures; the S_1 gains show that an informativeness–cost trade-off accounts for a systematic part of that choice.

The cost-informativeness trade-off

The ablations suggest a clear division of labor. Cost alone captures economy but can select cheap, uninformative fragments; semantics captures recoverability but can overgenerate long or redundant forms (see Tab. 1). The S_1 models combine these pressures, favoring candidates that remain meaning-aligned while avoiding costly or morphologically excessive alternatives. Figure 4 supports this interpretation. Semantic-only predictions are longest, consistent with maximizing semantic coverage without an economy term, while cost-only predictions remain short. The S_1 curves begin closer to cost-only and lengthen as k increases, suggesting that pragmatic ranking relaxes economy when extra morphological material helps preserve meaning.

The linear S_1 weights point the same way. Averaged over checkpoints, both the semantic weight and the coefficient on the cost-based score are positive ($\alpha = 0.38$, $\beta_{\text{score}} = 0.48$). Because the cost-based score is negated production cost, a positive β_{score} penalizes costly candidates. The ratio $\beta_{\text{score}}/\alpha \approx 1.28$ depends on normalization and the learned base models, but confirms that both pressures remain active in the fitted RSA utility.

This also explains why the S_1 advantage is most visible beyond the top ranks. At low k, highly salient semantic cues dominate; as the candidate set expands, many alternatives become plausible, and cost helps select among them. The pragmatic model is therefore most useful precisely when morphological choice is underdetermined by meaning alone.

Assumptions about speaker and listener knowledge

Our framework estimates RSA quantities from aggregate corpus statistics rather than individual judgments. Semantic compatibility is operationalized via distributional similarity, while production costs reflect population-level morpheme frequencies and form complexity. Thus $C(c,t)$ should be read as a historically available community-level candidate space, not as a set of alternatives explicitly considered by any single speaker.

This is a system-level use of RSA. Like information-theoretic efficiency accounts (Xu et al., 2024; Zaslavsky et al., 2018), it treats communicative utility as a pressure on the language system rather than a direct measurement of an individual. Its contribution is to make candidate comparison explicit: for a target concept at time t, which available morpheme sequence best balances listener recoverability and production cost? Future behavioral work can test whether speakers exhibit the same trade-off when coining novel morphological forms.

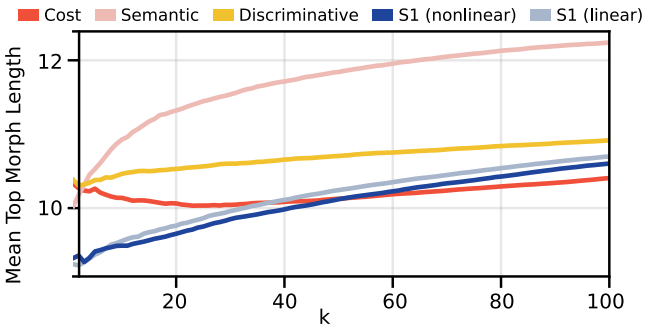

Figure 4: **Mean morphological length of top-k predictions vs. k.** The semantic baseline consistently generates the longest candidates; S_1 variants start shorter but grow steadily with k, eventually surpassing the cost-only model, reflecting a progressive relaxation of economy constraints to maintain semantic adequacy. Discriminative and cost-only models remain relatively stable.

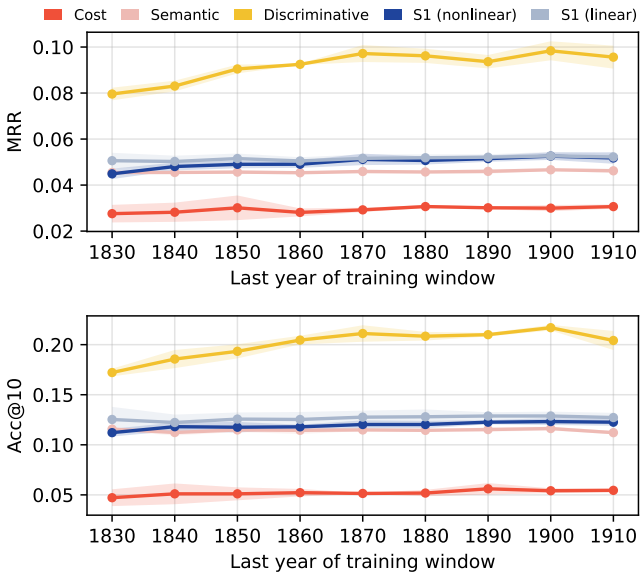

Figure 5: **Temporal robustness across cumulative training windows.** Ranking performance as a function of the last year included in the cumulative training window. MRR (left) and Acc@10 (right) show stable model ordering across windows; shaded bands denote ±1 standard deviation across folds.

Diachronic effects and the role of $\mathcal{L}_t$

Conditioning candidate availability and production cost on $\mathcal{L}_t$ gives the model a diachronic mechanism: as morpheme inventories, frequencies, and distributional neighborhoods change, the most efficient composition for a given concept may shift accordingly. To test whether the main ordering holds across the historical range, we retrained all model families on cumulative windows ending between 1830 and 1910 and evaluated each on its matched held-out set (Fig. 5). The broad ordering is stable: the discriminative model performs best, S_1 remains above the single-factor baselines, and cost-only remains lowest. Because data coverage varies across windows and semantic scoring relies on contemporary embeddings, this should be read as a robustness check rather than a full model of diachronic semantic change.

The main temporal limitation is that semantic scoring still relies on contemporary Qwen3 embeddings, so semantic drift is represented only indirectly through candidate generation and historical frequency features. Diachronic semantic encoders or historical definition sources would allow a more faithful reconstruction of listener expectations at time t.

The RSA–discriminative gap

The discriminative model's advantage reveals where the current RSA operationalization is too compressed. It has access to the full 13-dimensional feature vector and can learn interactions among semantic and cost features, whereas the S_1 models combine only two frozen scalar scores. This design preserves the interpretable RSA decomposition but discards cross-feature interactions, such as cases where a rare morpheme is acceptable because it is highly diagnostic for the target concept.

We therefore interpret the discriminative advantage as evidence that morphological composition depends on structure beyond a two-term utility. A natural next step is to enrich the RSA model without abandoning its decomposition, for example by jointly fine-tuning the semantic and cost components, incorporating morphotactic constraints, or allowing morpheme cost to depend on contextual informativeness.

Limitations and Future Directions

Morphological structure. Our model treats candidate morpheme sequences as unordered sets, ignoring morphotactic constraints, affix ordering, and head-modifier structure. Incorporating ordering constraints or sequence-sensitive composition functions would better reflect speakers' knowledge of morphological well-formedness and reduce structurally implausible candidates.

Historical and temporal grounding. Semantic compatibility is estimated using Qwen3 embeddings trained on contemporary corpora, and concept emergence is proxied by first attestation rather than by earlier conceptual emergence, which lexicalization may lag. Both approximations introduce uncertainty into $\mathcal{L}_t$ reconstruction, compounded by uneven corpus coverage across decades. Models such as diachronic embedding (Ma et al., 2025) would address these gaps.

Cross-linguistic generalization. We evaluate exclusively on English. Because compounding, affixation, and the informativeness–cost balance differ substantially across isolating, fusional, and agglutinative languages, extending the framework to typologically diverse languages is necessary to assess whether the communicative principles are universal.

Conclusion

We develop a computational account of morphological composition as rational communication, formalizing word formation as an RSA-style choice among competing morpheme sequences in a time-indexed lexicon $\mathcal{L}_t$. The central prediction is that languages prefer compositions that are both semantically recoverable for a listener and efficient for a speaker, capturing the classic informativeness–cost trade-off.

Ranking experiments on 4323 naturally occurring English compounds and derivations support this prediction. Models integrating semantic compatibility with production cost outperform single-factor baselines on the year-stratified held-out split, while the discriminative model marks the remaining predictive signal outside the constrained S_1 decomposition. Qualitative comparisons further show that pragmatic integration avoids the characteristic failure modes of single-factor approaches, prioritizing candidates that are meaning-aligned without becoming unnecessarily long or morphologically odd.

Taken together, our results suggest that morphological composition can be modeled as an emergent optimization shaped by the inventory and usage statistics of the historical lexicon, extending the rational communication program from utterance-level phenomena to the internal structure of words. Whether the same principles generalize across languages and word-formation types remains an open question.

Acknowledgement Fig. 1 was produced with AI-assisted illustration tools; all other figures were generated programmatically from experimental data. The authors would like to thank Guangyuan Jiang (MIT), Hongjie Li (PKU), Prof. Yanchao Bi (PKU), Prof. Huichao Yang (Hebei Normal University), and Dr. Qian Wang (PKU) for many valuable discussions and support. This work is supported in part by the National Natural Science Foundation of China (32595491, 62376009), the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

References

Affixes.org. (2008). Affixes.org: Dictionary of english affixes. (Cit. on p. 3).

Algeo, J. (1980). Where do all the new words come from? *American Speech*, *55*(4), 264–277 (cit. on p. 1).

Arndt-Lappe, S. (2015). Word-formation and analogy. In *Word-formation: An international handbook of the languages of europe* (pp. 822–841). De Gruyter Mouton. (Cit. on p. 2).

Bauer, L. (2001). *Morphological productivity* (Vol. 95). Cambridge University Press. (Cit. on pp. 1, 2).

Brinton, L. J., & Traugott, E. C. (2005). *Lexicalization and language change*. Cambridge University Press. (Cit. on pp. 1, 3).

Chandra, K., Chen, T., Li, T.-M., Ragan-Kelley, J., & Tenenbaum, J. (2024). Cooperative explanation as rational communication. *Annual Meeting of the Cognitive Science Society (CogSci)* (cit. on p. 2).

Chen, G., Van Deemter, K., & Lin, C. (2018). Modelling pro-drop with the rational speech acts model. *International Conference on Natural Language Generation* (cit. on p. 2).

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of english. *Literary and Linguistic Computing*, *25*(4), 447–464 (cit. on p. 3).

Davies, M. (2012). The 400 million word Corpus of Historical American English (1810–2009). In *English historical linguistics 2010* (pp. 231–262). John Benjamins Publishing Company. (Cit. on p. 3).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998 (cit. on pp. 1, 2).

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407 (cit. on pp. 1, 2).

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829 (cit. on pp. 1, 2).

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62 (cit. on pp. 1, 3).

Jiang, G., Hofer, M., Mao, J., Wong, L., Tenenbaum, J., & Levy, R. (2024). Finding structure in logographic writing with library learning. *Annual Meeting of the Cognitive Science Society (CogSci)* (cit. on p. 1).

Jiang, G., Hofer, M., Mao, J., Wong, L., Tenenbaum, J. B., & Levy, R. (2025). Finding structure in logographic writing with library learning ii: Grapheme, sound, and meaning systematicity. *Annual Meeting of the Cognitive Science Society (CogSci)* (cit. on p. 1).

Levin, B., Glass, L., & Jurafsky, D. (2019). Systematicity in the semantics of noun compounds: The role of artifacts vs. natural kinds. *Linguistics*, *57*(3), 429–471 (cit. on p. 1).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177 (cit. on pp. 1, 3).

Ma, Y., Peng, Y., & Zhu, Y. (2025). Word embeddings track social group changes across 70 dates in china. *Annual Meeting of the Cognitive Science Society (CogSci)* (cit. on p. 6).

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, *122*(3), 485–515 (cit. on p. 1).

Mattiello, E., & Dressler, W. U. (2018). The morphosemantic transparency/opacity of novel english analogical compounds and compound families. *Studia Anglica Posnaniensia*, *53*(1), 67–114 (cit. on p. 1).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (cit. on p. 3).

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM*, *38*(11), 39–41 (cit. on pp. 2, 3).

Peng, Y., Ma, Y., Wang, M., Wang, Y., Wang, Y., Zhang, C., Zhu, Y., & Zheng, Z. (2025). Probing and inducing combinational creativity in vision-language models. *Annual Meeting of the Cognitive Science Society (CogSci)* (cit. on p. 1).

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences (PNAS)*, *108*(9), 3526–3529 (cit. on p. 1).

Plag, I. (1999). *Morphological productivity: Structural constraints in english derivation* (Vol. 28). Walter de Gruyter. (Cit. on pp. 1, 2).

Reddy, S., McCarthy, D., & Manandhar, S. (2011). An empirical study on compositionality in compound nouns. *International Joint Conference on Natural Language Processing* (cit. on p. 1).

Salge, C., Ay, N., Polani, D., & Prokopenko, M. (2015). Zipf's law: Balancing signal usage cost and communication efficiency. *PLOS One*, *10*(10), e0139475 (cit. on p. 1).

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423 (cit. on p. 1).

Štekauer, P., & Lieber, R. (Eds.). (2005). *Handbook of word-formation* (Vol. 64). Springer Science & Business Media. (Cit. on p. 1).

TheWelcomer. (2025). *MorphSeg* [GitHub repository]. https://github.com/TheWelcomer/MorphSeg (cit. on p. 3).

Xu, A., Kemp, C., Frermann, L., & Xu, Y. (2024). Word reuse and combination support efficient communication of emerging concepts. *Proceedings of the National Academy of Sciences (PNAS)*, *121*(46), e2406971121 (cit. on pp. 1, 2, 5).

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences (PNAS)*, *115*(31), 7937–7942 (cit. on pp. 2, 5).

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176* (cit. on p. 3).

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press. (Cit. on p. 1).