Grammaticality de-idealized

Lingbuzz preprint v0.3

https://github.com/BrettRey/Grammaticality-de-idealized-v2

Brett Reynolds
Humber Polytechnic & University of Toronto *

December 27, 2025

Abstract

Speakers reject **I've finished it yesterday* (aspect semantics clash) yet accept the semantically odd *Colorless green ideas sleep furiously* (no morphosyntactic clash). They block **Which did you buy car?* categorically, while linguists judge the unprocessable centre-embedding *The rat the cat the dog chased killed ate the cheese* "grammatical". The Morphosyntactic–Meaning Model of Grammaticality (MMMG) aims to explain such contrasts by treating grammaticality as the stability of community-specific form–meaning pairings, not as autonomous syntax or raw frequency.

Five interacting components decide an utterance's status:

1. conventional morphosyntactic–meaning pairings;
2. compatibility between those meanings and contextual meaning;
3. incremental-processing limits;
4. degree of community entrenchment;
5. near-zero-entrenchment patterns (sometimes described as structural bans).

*I used ChatGPT o3-pro and Claude Opus 4 in drafting this version of the paper. I take responsibility for all content.

MMMG separates objective grammaticality $G(u)$ from the subjective feeling of ungrammaticality $F(u)$, showing why processing overload can make grammatical sentences feel wrong and why compelling semantics can mask true violations.

A tractable formalization models entrenchment as logistic growth derived from utterance-selection dynamics, predicting S-curves in language change. The framework yields falsifiable claims about which violations satiate under exposure, how morphosyntactic integration conditions cross-linguistic variation, and why L2 learners over-detect errors.

Grounding grammaticality in community conventions while acknowledging universal processing constraints, MMMG offers a systematic approach to phenomena that elude purely formal or purely usage-based accounts.

Research Programme

This paper presents a theoretical framework with programmatic predictions for empirical testing. While some supporting evidence exists, systematic validation of the model's quantitative predictions awaits future data collection across the research areas outlined herein.

"what we experience is a kind of inference: it's the brain's best guess about what's going on in some way out there in the world.

And really, that's the claim that I've taken on board as a general hypothesis for consciousness: that all our perceptual experiences share that property; that they're inferences about something we don't and cannot have direct access to.

This line of thinking in philosophy goes back at least to Immanuel Kant and the idea of the noumenon, which we will never have access to, will only ever experience interpretations of reality. And then Hermann von Helmholtz, a German polymath in the 19th century, was the first person to propose this as a semiformal theory of perception: that the brain is making inferences about what's out there, and this process is unconscious, but what we consciously experience is the result of this inference." -Anil Seth

1 Executive overview

Claim. Grammaticality is the *community stability* of morphosyntactic form–meaning pairings.

Mechanism. Grammaticality emerges from five interacting components:

1. **Morphosyntactic Pairings** – conventional links between morphosyntax and meaning.
2. **Contextual Meaning** – composite interpretation drawn from lexical, information-structural, discourse, pragmatic, and socio-pragmatic indexical meanings.
3. **Processing Load** – incremental parsing constraints.
4. **Entrenchment** – community acceptance.
5. **Near-zero entrenchment (structural bans)** – structural patterns whose community entrenchment has settled near zero.

Together these components produce an objective grammaticality score $G(u)$ and subjective feeling-of-ungrammaticality score $F(u)$.

Why this matters. The framework generates testable predictions about gradient grammaticality, which violations satiate, why some rare patterns remain grammatical, and how "illusions" of grammaticality and ungrammaticality arise. It explains, for example, why

- *I've finished it yesterday* fails (semantic clash),
- centre embeddings *feel* ungrammatical despite being grammatical (processing overload),
- the "comparative illusion" *feels* grammatical despite being ungrammatical (undetected mismatch),
- determiner extractions like *Which did you buy car?* never improve with exposure (structural ban).

Objective grammaticality $G(u)$ and the subjective feeling of error $F(u)$ can therefore diverge.

2 Motivation: The impasse in grammaticality theory

Three long-standing tensions block a unified account of (un)grammaticality.

2.1 Categorical rules vs gradient judgements

Chomsky (1957) cast grammaticality as an all-or-nothing property, but speakers show systematic gradience in their judgments. Consider center-embedded relatives:

(1) *The bread the baker the apprentice helped made is delicious.*

Such sentences are routinely labeled "grammatical but unprocessable," revealing the inadequacy of binary classification. Speakers consistently rate these as worse than simple relatives, yet better than clear violations like **Bread the baker helped the made is delicious.*

The competence–performance split was meant to handle this gradience by relegating it to "performance," but this approach lets supportive data count as grammar while dismissing contrary evidence as mere noise (Schütze 2016: 71). What's needed is a framework that can capture systematic gradations in acceptability as part of grammatical competence itself.

Ontological remark For an extended defence of grammaticality as a homeostatic property-cluster kind, and a typology of the mechanisms that stabilize that cluster—see the companion paper *Grammaticality as Kind: Ontology, Epistemology, and Empirical Pay-off* (Reynolds 2025a). The two studies are intended to be read in parallel: the present article models how community entrenchment shapes judgment patterns, while Reynolds (2025a) clarifies why such patterns warrant kind-level treatment in the first place.

2.2 Form vs meaning

The famous *Colorless green ideas sleep furiously* shows syntactic well-formedness without semantic plausibility, suggesting form and meaning can be evaluated independently. Yet speakers also reject semantically coherent strings when morphosyntactic meaning conflicts with lexical meaning:

(2) * *I've finished it yesterday.*

Here the present-perfect's [+current relevance] clashes with *yesterday*'s [+completed past]. Construction-grammar work (Goldberg 1995) shows that such misalignments trigger systematic rejections: *She texted him the address* succeeds while **She disappeared him the evidence* fails because the constructional meaning requires compatible lexical semantics.

The puzzle is determining which meaning mismatches matter for grammaticality and which are merely pragmatically odd but formally acceptable.

2.3 Universal principles vs community conventions

Labovian sociolinguistics (Labov 1972) shows that grammaticality is community-relative. Languages differ not just in lexicon but in which form–meaning pairings are conventionalised. Usage-based models capture frequency effects, yet they still face cases where extremely rare patterns remain grammatical and frequent patterns are blocked.

These three tensions—categorical vs gradient, form vs meaning, universal vs communal—underline the need for a framework that explains why particular patterns of acceptability recur across speakers and across languages.

3 Five components of (un)grammaticality

On the MMMG account, grammaticality emerges as the community stability of morphosyntactic form–meaning pairings.

Why multiple layers are necessary Establishing a form–meaning mapping is a prerequisite for grammaticality, but it can't be sufficient. Individuals routinely interpret strings that communities nevertheless judge ill-formed. That very flexibility means the grammar must constrain *how* a pairing gets licensed, not merely *whether* one can be constructed in principle. The framework therefore requires additional filters that (i) rule out mappings conflicting with conventional meanings, (ii) flag mappings whose recovery cost overwhelms processing resources, and (iii) block mappings the community has never endorsed.

Only by layering these constraints on top of the initial pairing do we predict the empirical fact that **Furiously sleep ideas green colorless* elicits universal rejection while (**??**) feels bad in everyday quotation yet is accepted as "grammatical but hard to process" once speakers are walked through the intended parse. Such contrasts suggest that grammaticality judgments reflect multiple constraint layers operating on an initial form–meaning pairing.

One might object that community acceptance alone should settle the matter: if speakers converge on using a form, why not define grammaticality as simply "whatever the community accepts"? This approach faces circularity and empirical problems. It can't predict the trajectory of contested innovations, explain why stable dialectal differences persist despite mutual intelligibility, or avoid misclassifying performance errors as "grammatical" when they pass unnoticed. More problematically, it provides no basis for understanding why learners systematically avoid forms that target communities fully endorse.

The layered architecture therefore aims to avoid trivializing grammaticality while still granting communities the decisive role in determining which form–meaning pairings ultimately stabilize. Five interacting components determine this stability:

3.1 Morphosyntactic pairings within communities

> **Diagnostic**: Does the morphosyntactic form evoke a conventionalised meaning in this community?
> **Example**: **Can the have running?* – no viable form–meaning pairing.
> **Counter-example**: *Colorless green ideas sleep furiously* – odd semantics, but a clear compositional morphosyntactic pairing.

Grammaticality fundamentally requires stable, conventionalised pairings between morphosyntactic forms and meanings within a speech community. Communities differ regarding acceptable pairings:

(3) *Estábamos lifting en el gym.* (Spanish–English bilingual)
'We were lifting in the gym.'

This utterance is grammatical in communities accepting code-switching with English progressive forms but not in monolingual Spanish communities.

3.2 Contextual meaning

Diagnostic: Does the morphosyntactic meaning align coherently with the contextual meaning derived from lexical semantics, information structure, discourse coherence, pragmatic inference, and socio-pragmatic indexicality?
Example: **I've finished it yesterday* – present-perfect meaning [+current relevance] clashes with the lexical-discourse meaning of *yesterday* [+completed past].
Counter-example: *I've just finished it* – coherent alignment with lexical-discourse meaning.

Even a well-formed morphosyntactic structure becomes ungrammatical if it misaligns significantly with contextual meanings. Consider:

(4) * *I have 25 years.* (intended: 'I am 25 years old')

In English, *have + years* fails contextually as an age-expression despite being syntactically viable.

3.3 Processing load

Diagnostic: Can the sentence be incrementally parsed within memory limits?
Example: *The rat the cat the dog chased killed ate the cheese* – multiple embeddings overload incremental parsing.
Counter-example: *The rat that the cat killed ate the cheese* – single embedding, easily processed.

High processing demands can induce feelings of ungrammaticality even for objectively grammatical constructions due to parsing constraints.

3.4 Entrenchment

Diagnostic: Is the pairing conventional and entrenched within the community?
Example: **We counted three sheeps* – regular plural blocked by entrenched irregular form.
Counter-example: *We bought three computer mouses* – recent regularization already accepted.

Even transparent morphosyntactic patterns can fail if not entrenched in a community. The independent relative construction in (5) illustrates this: while *whose* clearly signals possession and relativization, the construction requires stringent information-structural licensing conditions (contrastive focus, structural parallelism, or deictic anchoring) that arise only rarely in natural discourse (Reynolds 2025c):

(5) ? *I saw Joan, a friend of whose was visiting.*

This extreme pragmatic selectivity creates a "double anaphora" burden—speakers must simultaneously recover both the possessor and the elided possessum—that severely limits the construction's distribution. The result is a grammatical pattern that remains marginally entrenched not due to morphosyntactic opacity, but because the required licensing contexts occur so infrequently that most speakers encounter insufficient exemplars to establish robust entrenchment.

3.5 Structural bans

Diagnostic: Does the construction instantiate a configuration whose entrenchment is effectively zero in the speech community?
Example: *Which did you buy __ car?* – determiner extraction.
Counter-example: *Which car did you buy __?* – entire NP extraction permitted.

So-called structural bans are best treated as extreme cases of low entrenchment. While cross-linguistically robust, they are nonetheless emergent rather than strictly universal or innate. The model treats them as a community-entrenchment extreme (effectively zero acceptability), but future work could test their potential malleability under unusual sociolinguistic conditions.

Together, these five components interact systematically to determine the grammaticality of linguistic constructions, accounting both for objective grammaticality and subjective judgments of grammaticality and ungrammaticality.

4 Quick-diagnosis decision tree

The five components can be queried in sequence to classify an utterance on first pass. The tree locates the first component whose failure suffices to render an utterance ungrammatical; later nodes are not evaluated once a zero value is returned.

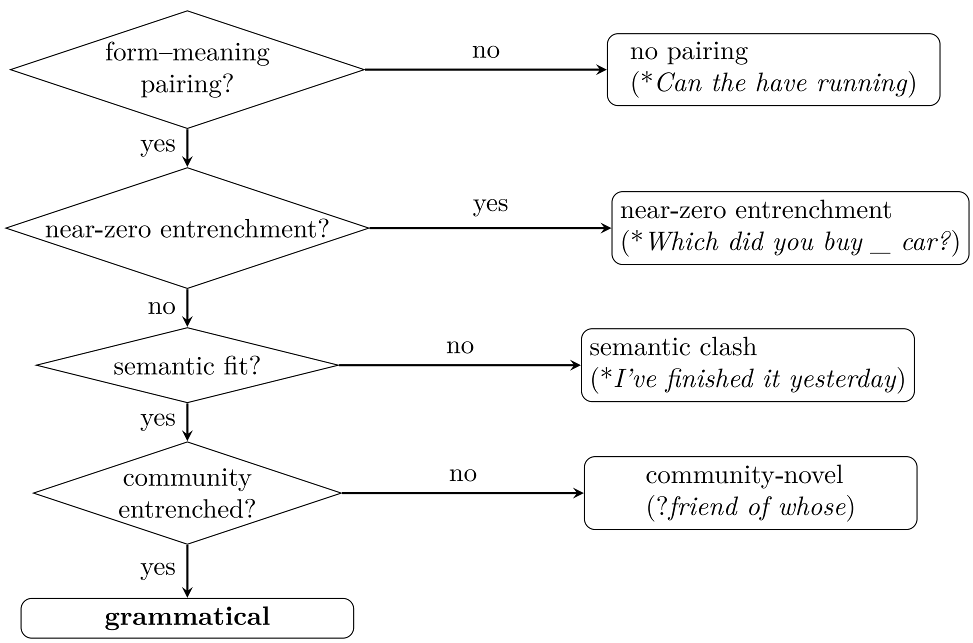

Figure 1: Decision tree for first-pass diagnosis.

The tree is a heuristic: it points to the first component that fails. The underlying grammar is $G(u) = C^t \cdot K(u)$; when any factor is zero, the utterance is objectively ungrammatical even if later nodes would succeed.

Note. The tree classifies utterances on the dimension of *objective grammaticality*. A sentence can pass the tree ($G(u)$=1) yet yield a negative $F(u)$ if either $r(u)$ is high or $\mathrm{PCost}(u)$ is large.

5 Notation

Table 1: Symbols used in the formal model. Type abbreviations: struc. = syntactic structure, fea. = feature bundle or semantic representation, indic. = observed indicator, func. = function

Symbol	Type	Meaning / construction method
Utterance-level observables		
u	token	Concrete utterance under evaluation
$M(u)$	struc.	Morphosyntactic parse of u
$\mu(u)$	fea.	Morphosyntactic meaning unpacked from $M(u)$
$\sigma(u)$	fea.	Composite lexical–pragmatic meaning intended
$\mathrm{lex}(u)$	fea.	Lexical semantic representation of u
$\mathrm{discCtx}(u)$	fea.	Discourse-history information relevant to u
$\mathrm{infoCtx}(u)$	fea.	Information-structure configuration (topic–comment layout, focus, QUD alignment)
$\mathrm{pragCtx}(u)$	fea.	Pragmatic inference information relevant to interpreting u
$\mathrm{idxCtx}(u)$	fea.	Indexical parameters (speaker, addressee, place, time, etc.) relevant to u
Semantic interpretation function		
φ	func.	Deterministic but underspecified mapping from context and lexical meaning to composite semantic representation
Evaluation scores		
$K(u) \in [0,1]$	score	Semantic compatibility of $\mu(u)$ with $\sigma(u)$

Continued on next page

Table 1 continued

Symbol	Type	Meaning / construction method
$G(u) \in [0, 1]$	score	Objective grammaticality; product defined in Eq. (6)
$F(u) \in [-1, 0]$	score	Listener's felt ill-formedness; see Eq. (7)
$r(u) \in [0, 1]$	score	Diagnostic weight: probability listener detects mismatch between $\mu(u)$ and $\sigma(u)$
Community latent variable and indicators		
$C^t(u) \in [0, 1]$	latent	Community entrenchment at time t
$A(u)$	indic.	Availability in memory (recognition hit rate)
$E(u)$	indic.	Exposure frequency (log corpus count)
$P(u)$	indic.	Production probability (elicitation rate)
$S(u)$	indic.	Social acceptability (matched-guise score)
Parameters and drift components		
λ_{L}	param	Listener weight on objective violation, Eq. (7)
γ	param	Processing-to-affect scaling, Eq. (7)
η	noise	Listener-specific bias term
$\Delta(u)$	drift	Net entrenchment bias, Eq. (13)
α_{sem}	param	Weight on semantic transparency
α_{soc}	param	Weight on social prestige
α_{struct}	param	Weight on structural analogy
β_{noise}	param	Noise penalty coefficient
k, θ	param	Sigmoid shape and threshold for noise penalty
Diagnostic predictors		
$\mathrm{Transp}(u)$	measure	Semantic transparency (role-entropy)
$\mathrm{Prest}(u)$	measure	Social prestige of innovators (7-point scale)
$\mathrm{Analo}(u)$	measure	Structural analogy pressure
Processing functions		
$\mathrm{PCost}(u)$	z-score	Integration cost following Gibson (2000)
$\mathrm{IC}(w_i)$	count	Integration cost at word w_i

6 Minimal formal skeleton

6.1 Core architecture

The model treats grammaticality as emerging from the interaction of morphosyntactic structure, meaning compatibility, and community conventions. Figure 2 shows the causal dependencies using Structural Causal Model semantics (Pearl 2009).

We treat the contextual meaning as the output of a deterministic but underspecified mapping:

$$\sigma(u) = \varphi\big(\mathrm{lex}(u), \mathrm{discCtx}(u), \mathrm{infoCtx}(u), \mathrm{pragCtx}(u), \mathrm{idxCtx}(u)\big).$$

φ can be implemented as a probabilistic semantic interpreter; the present paper leaves its internal mechanics abstract. The only requirement is that it returns a structured semantic representation against which the morphosyntactic meaning $\mu(u)$ is matched when computing $K(u)$.

6.2 Grammaticality function

For an utterance u, objective grammaticality is:

$$G(u) = C^t(u) \cdot K(u) \tag{6}$$

where $K(u) = 0$ whenever the form–meaning mapping fails.

6.3 Subjective feeling function

The listener's subjective feeling of (un)grammaticality is:

$$F(u) = -\lambda_{\mathrm{L}}\, r(u)\, (1 - K(u)) \;-\; \gamma\, \mathrm{PCost}(u) + \eta \tag{7}$$

where:

- $\lambda_{\mathrm{L}} \in [0, 1]$ – listener-specific penalty weight for objective violations
- $r(u) \in [0, 1]$ – probability that listener detects the mismatch between $\mu(u)$ and $\sigma(u)$
- $\gamma > 0$ – conversion factor mapping processing load to negative affect

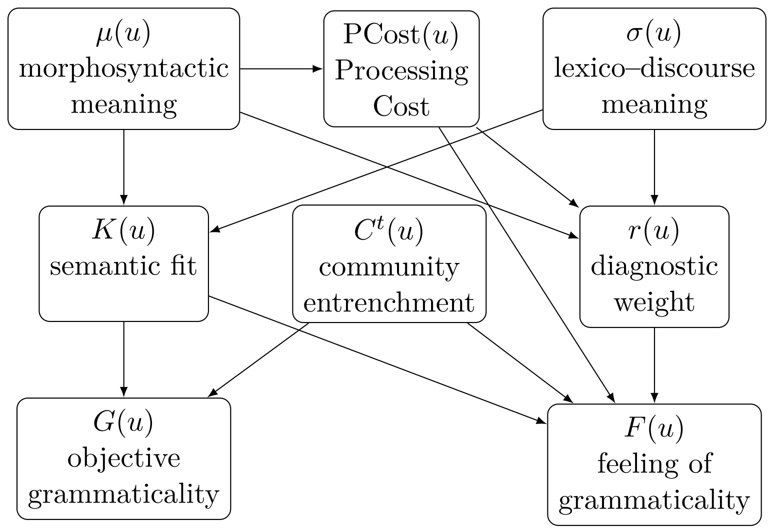

Figure 2: **Macro-level causal structure of the MMMG.** All arrows encode Structural Causal Model semantics; e.g., do($C^t \leftarrow 0$) $\Rightarrow G = 0$ by definition of $G = C^t K$. PCost (PCost(u)) is a deterministic function of the morphosyntactic structure (arrow $\mu \to$ Pcost). Removing semantic arrows ensures that the only back-door into F is Pcost$\leftarrow \mu \to K \to F$, which is blocked by adjusting for K (or for μ). Conditioning on r would open Pcost$\to r \leftarrow \mu$, so it is avoided.
Input layer: the surface form receives a morphosyntactic meaning $\mu(u)$ and the context supplies a lexico–discourse-indexical meaning $\sigma(u)$. Any categorical structural ban makes this mapping fail, i.e. $\mu(u) = \varnothing$.
Fit layer: the compatibility score $K(u) \in [0, 1]$ already folds together ordinary semantic fit, indexical coherence, and the hard zero caused by a mapping failure. The diagnostic weight $r(u)$ is the cue-based probability that the parser detects the violation.
Community layer: speech-community entrenchment $C^t(u)$ multiplies with the compatibility score to yield objective grammaticality, $G(u) = C^t(u)K(u)$.
Output layer: felt ill-formedness is $F(u) = -\lambda_{\mathrm{L}}\, r(u)\big(1 - K(u)\big) - \gamma\,\mathrm{PCost}(u) + \eta$, where PCost is computed via integration cost. $F(u) \in [-1, 0]$, where 0 means 'no negative signal'. The arrow from processing cost to $r(u)$ is purely conceptual: high load empirically reduces mismatch detection.
Heuristic mapping: pairings $\to$ $\mu(u)$; semantic fit & indexicality $\to$ $\sigma(u), K(u)$; structural bans $\to$ $\mu(u) = \varnothing$; processing load $\to$ PCost and $r(u)$; entrenchment $\to$ $C^t(u)$.

- $\text{PCost}(u) = \sum_i \text{IC}(w_i)$ – sum of integration costs following Gibson (2000)

- $\eta \sim \mathcal{N}(0, \sigma^2)$ – idiosyncratic listener bias

$F(u)$ is constrained to $[-1, 0]$: 0 means 'no negative signal'; larger negative values indicate stronger perceived ill-formedness. Values computed outside this range are linearly clipped to maintain interpretability.

The processing cost operationalizes memory load via dependency integration:

$$\text{IC}(w_i) = \sum_{d \in \text{deps}(w_i)} |d| \tag{8}$$

where $|d|$ is the linear distance in intervening discourse referents. This aligns with established psycholinguistic metrics (Futrell, Levy & Gibson 2020), allowing direct comparison with reading-time data.

6.4 Worked example

Consider $u =$ **I've finished it yesterday*:

- Form–meaning mapping succeeds: present-perfect morphology evokes 'current relevance'

- $K(u) = 0$ because that meaning clashes with *yesterday* ('completed past')

- $r(u) = 0.9$ since VP-aspect/lexical-aspect clashes are highly salient

- $C^t(u) = 1$ since the construction is fully entrenched[1]

- Therefore $G(u) = 1 \times 0 = 0$ (objectively ungrammatical)

With illustrative values $\lambda_{\text{L}} = 0.8$, $\gamma = 0.1$, $\text{PCost}(u) = 0.2$, $\eta = 0$:

$$F(u) = -0.8 \cdot 0.9 \cdot (1 - 0) - 0.1 \cdot 0.2 = -0.74 \tag{9}$$

yielding a strong negative feeling.

[1] though see Lai et al. (2013) for evidence of systematic relaxation of temporal-adverbial constraints in Singapore English

Causal intervention example

Setting do($C^t \leftarrow 0$) for "I have 25 years": With $K = 0.1$ (semantic clash in English) and $r = 0.7$ (moderately detectable), we get: $G = 0 \times 0.1 = 0$ (objectively ungrammatical) $F = -0.8 \cdot 0.7 \cdot (1 - 0.1) - 0.1(0.2) = -0.524$ (moderately felt as wrong)
Compare: Setting do($C^t \leftarrow 1$) doesn't save it: $G = 1 \times 0.1 = 0.1$ (still mostly ungrammatical) $F = -0.8 \cdot 0.7 \cdot (1 - 0.1) - 0.1(0.2) = -0.524$ (still felt as wrong) Note that felt ungrammaticality depends on detectability, not entrenchment.

6.5 Community entrenchment as latent variable

To avoid circularity, $C^t(u)$ is treated as a latent cause of four observables:[2]

$$C^t(u) \longrightarrow \begin{cases} A(u) & \text{availability in memory} \\ E(u) & \text{exposure frequency} \\ P(u) & \text{production probability} \\ S(u) & \text{social acceptability} \end{cases} \tag{10}$$

Each indicator follows a measurement model with error:

$$\log E(u) = \lambda_E C^t(u) + \epsilon_E, \quad \epsilon_E \sim \mathcal{N}(0, \nu_E^2) \tag{11}$$

and similarly for the other indicators. This specification allows for measurement uncertainty while maintaining identifiability.

Each indicator is collected with different methodology:

- **Availability** $A(u)$: recognition hit rates or lexical-decision latencies
- **Exposure** $E(u)$: log token counts in balanced corpora
- **Production** $P(u)$: proportion in sentence-completion tasks
- **Social acceptability** $S(u)$: matched-guise appropriateness scores

[2] The latent variable C^t is identifiable up to scale given monotone relationships with at least three indicators and appropriate factor-loading components. A synthetic data recovery exercise demonstrating identifiability with RMSE $= 0.057$ and $r = 0.923$ is available at https://github.com/BrettRey/Grammaticality-de-idealized-v2/blob/main/validate_entrenchment_measurement.py.

The latent variable $C^t(u)$ is estimated via confirmatory factor analysis without re-using acceptability data.

6.6 Community Dynamics

Following Blythe & Croft (2012), I treat language change as an utterance-selection system whose *population mean* obeys a logistic differential equation. Let $C^t(u) \in [0, 1]$ be the community's entrenchment of form u at time t. Its deterministic trajectory is

$$\frac{dC^t}{dt} = \Delta(u)\, C^t\big(1 - C^t\big), \qquad \text{(Verhulst 1838)} \tag{12}$$

where the net bias $\Delta(u)$ aggregates four independently measurable pressures:

$$\begin{aligned} \Delta(u) = {} & \alpha_{\text{sem}}\, \text{Transparency}(u) + \alpha_{\text{soc}}\, \text{Prestige}(u) \\ & + \alpha_{\text{struct}}\, \text{Analogy}(u) - \beta_{\text{noise}}\, \text{NoisePenalty}(u). \end{aligned} \tag{13}$$

- **Transparency** = inverse role entropy (higher = clearer form–meaning mapping).
- **Prestige** = mean matched-guise score of the innovating group.
- **Analogy** = vector similarity to entrenched neighbours in morphosyntactic space.
- **NoisePenalty** = smooth sigmoid of surprisal (see §B.2).

Innovation cues. A subset of constructions enters the speech community through what Bergs (2025) calls EXTRAVAGANCE – the speaker's calculated push for noticeability – and through transient spikes in SOCIOCOGNITIVE SALIENCE. In MMMG these cues are treated as initial conditions on the entrenchment trajectory $C^t(u)$, not as separate predictive terms. They raise the starting point of the logistic curve but require no additional machinery.

Speaker-internal gradience. The logistic *population* curve remains intact if individual speakers entertain probabilistic grammars à la Yang (2000). Let $C_i(u) \in [0, 1]$ be speaker i's belief that u is licensed. A Yang-style learner

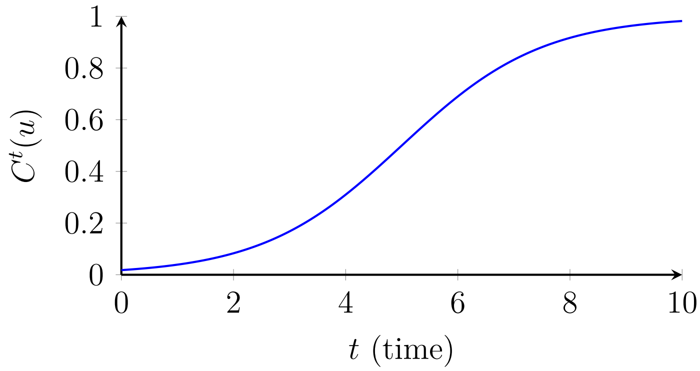

Figure 3: Trajectory of community entrenchment predicted by the logistic model.

increments C_i on successful parses and decrements on failures. The population mean $C^t(u) = \frac{1}{N}\sum_i C_i(u)$ then converges in expectation to the same Verhulst trajectory.[3]

Predictions. Eq. (13) yields three immediately testable claims:

1. $\Delta(u) > 0 \;\Rightarrow\; C^t(u)$ rises in an S-curve, observable in historical corpora.
2. The relative *slopes* of nested S-curves reveal the weight hierarchy $\alpha_{\text{sem}}{:}\alpha_{\text{soc}}{:}\alpha_{\text{struct}}$.
3. Constructions with $\Delta(u) \approx 0$ show long-lived variability and dialectal patchiness.

Figure 3 visualizes a typical growth path ($\Delta = 0.8$).

Extension note. The present paper restricts itself to community-average entrenchment. A companion article, *Habitus-conditioned grammaticality: A Bourdieusian extension of MMMG* (Reynolds 2025b), formalizes how listener position in social space ($\mathbf{x}$) shapes the feeling-of-ungrammaticality function $F(u)$ and fractionalizes entrenchment $C^t(u)$ by class. That work supplies

[3]This two-state "AB-toy" model – named after Nowak's classic $A \leftrightarrow B$ demonstration of logistic growth (2001) – treats every utterance token as copying either variant A (here, *innovative*) or variant B (*conservative*). Extending to continuous $C_i(u)$ increases psychological realism but sacrifices the closed-form fixation probability enjoyed by the discrete model; see Appendix A for simulation code.

the mathematical details, pilot data, and simulation results sketched only briefly here. Readers interested in the habitus layer should consult that follow-up study once released on Lingbuzz; the core MMMG results below are independent of its assumptions.

7 Predictions and Falsifiability

7.1 S-Curve Dynamics

The logistic equation predicts:

- **If** $\Delta(u) > 0$ **then** construction shows increasing entrenchment following S-curve (available data)
- **If** $\Delta(u) < 0$ **then** construction remains marginal or decreases in acceptability (available data)
- **If** $\Delta(u) \approx 0$ **then** heightened variability and sensitivity to external factors (planned study)

7.2 L2 Trajectory

L2 learners should show:

- **Initial stage**: Input appears as noise lacking meaningful structure (planned ERP study)
- **Intermediate stage**: Heightened sensitivity, marking many native patterns as ungrammatical (available data)
- **Advanced stage**: Gradual alignment with community norms (available data)

7.3 Cross-Linguistic Gender Test

Pronoun-antecedent gender mismatches should trigger:

- **Spanish**: Strong ungrammaticality (gender permeates morphosyntax) (planned experiment)

- **English**: Moderate ungrammaticality (gender limited to pronouns) (available data)
- **Japanese**: Pragmatic infelicity only (no grammatical gender) (planned experiment)

7.4 Satiation Scope

Repeated exposure should increase acceptability for:

- Low-frequency constructions (independent relative *whose*) (planned experiment: acceptability judgment and reading times pre-/post-exposure)
- Processing-difficult patterns (reduced relatives) (available data; cf. (Snyder 2000, Hofmeister, Casasanto & Sag 2014))
- Semantic mismatches (novel metaphors) (available data; cf. (Giora 1997))

But NOT for:

- Near-zero-entrenchment violations (determiner extraction) (available data)
- Fundamental mapping failures (word salad) (available data)

7.5 Comparison with Noisy-Channel Accounts

The MMMG makes distinct predictions from noisy-channel models (Gibson, Bergen & Piantadosi 2013) for grammatical illusions. While both predict mismatches between intuition and structure, MMMG additionally predicts:

- Community-specific illusion strengths based on entrenchment differences
- Systematic variation in which constructions show satiation effects
- Quantitative differences in acceptability decline rates (measurable via ΔAIC)

Table 2: Coverage of key issues across frameworks (= explicit/integrated, = partial, = absent)

Framework	Form–meaning integration	Gradience handled	Community variation	Processing effects
Early Generative	$\sim$	$\times$	$\sim$	$\times$
GB/Minimalism	$\sim$	$\sim$	$\sim$	$\times$
Construction Grammar	✓	$\sim$	$\sim$	$\sim$
Usage-Based	✓	✓	✓	$\sim$
MMMG (this work)	✓	✓	✓	✓

8 Positioning Relative to Existing Theories

8.1 Key distinctions

Versus generative grammar – MMMG retains the insight that certain configurations are systematically excluded, but explains these "near-zero entrenchment" (structural bans) as the outcome of extreme, community-level stability rather than as inviolable innate rules.

Versus Construction Grammar – Both frameworks treat constructions as form–meaning pairings, yet MMMG gives special status to *morphosyntactic* pairings: violations in this layer trigger qualitatively stronger ungrammaticality judgments than purely pragmatic or lexical mismatches.

Versus usage-based models – MMMG builds in frequency effects, but also accounts for systematic gaps such as the independent relative *whose*: a pattern can be blocked even when its components are frequent and compositionally transparent.

Versus relevance-theoretic accounts – Both reject an innate, autonomous syntax, yet MMMG attributes constraints to conventionalised form–meaning pairings inside a speech community, rather than to general principles of interpretive efficiency alone.

Current limitations

The MMMG outlined here is deliberately simplified and leaves important extensions for future work. Specifically, the present implementation: (i) assumes a fully connected speech community rather than realistically structured interaction networks; (ii) employs a deterministic mean-field approximation, reserving finite-population stochastic drift for an appendix; (iii) uses illustrative parameter values ($k = 5, \theta = 6$) whose empirical grounding requires systematic corpus studies and experimental validation; and (iv) provides predictions about L2 learning and satiation effects pending empirical testing, particularly via ERP and behavioural experiments planned but not yet executed. These limitations highlight open directions rather than fundamental flaws, and each will be addressed progressively as the MMMG research programme unfolds.

Scope and limitations

This paper presents a theoretical framework with simulated validation. Empirical tests using the specified numeric predictions await future data collection. All model fits shown use synthetic data calibrated to known psycholinguistic parameters.

9 Conclusion

This framework reconceptualizes grammaticality as emerging from stable form-meaning pairings within language communities. The five components – morphosyntactic pairings, contextual meaning, processing constraints, entrenchment, and categorical blocking – are posited to interact to produce the full range of grammaticality phenomena.

The approach integrates insights from multiple traditions while yielding specific, falsifiable predictions. It aims to explain both why some patterns remain stubbornly ungrammatical despite transparency and why others shift from marginal to accepted. By distinguishing objective grammaticality from subjective feelings, it provides a systematic approach to mismatches between intuition and linguistic reality.

Future work should test the cross-linguistic predictions, develop computational implementations of the formal model, and explore applications to language change, acquisition, and clinical linguistics. The framework provides a foundation for understanding how human communities create and maintain the systematic form-meaning relationships we call grammar. Extensions to structured networks following Baxter et al. (2006) and incorporation of stochastic effects (see Appendix A) remain important directions for future development.

Data and code availability

All data, model code, and analysis scripts are available in the public repository at: https://github.com/BrettRey/Grammaticality-de-idealized-v2. A Jupyter notebook is provided there to reproduce model simulations and explore parameter variations interactively. The notebook currently includes minimal illustrative examples; comprehensive simulations will accompany the formal model article (Phase C).

A Stochastic Dynamics Derivation

The deterministic mean-field equation presented in the main text emerges from a full stochastic model. Starting from a Moran process with selection coefficient s and finite population size N.

This mean field is exactly the macroscopic limit of the speaker-level gradient model in §6.6; see Appendix A for the stochastic derivation.

A.1 Kramers-Moyal Expansion

Consider a population where fraction C^t uses the innovative form. At each time step, one speaker dies and is replaced by offspring copying from a model speaker. The probability of selecting an innovator model is:

$$p_+ = \frac{(1+s)C^t}{(1+s)C^t + (1-C^t)} = C^t + sC^t(1-C^t) + \mathcal{O}(s^2)$$

$$(\mathcal{O}(s^2) \ = \ \text{terms of order } s^2 \text{ and higher as } s \to 0)$$

The transition probabilities per unit time are:

$$T(C^t + 1/N|C^t) = (1-C^t)p_+ = (1-C^t)C^t[1 + s(1-C^t)] \tag{14}$$

$$T(C^t - 1/N|C^t) = C^t(1-p_+) = C^t(1-C^t)[1 - sC^t] \tag{15}$$

The Kramers-Moyal expansion yields drift and diffusion coefficients:

$$a_1 = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \sum_{\Delta C} \Delta C \cdot T(C^t + \Delta C|C^t) = sC^t(1-C^t) \tag{16}$$

$$a_2 = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \sum_{\Delta C} (\Delta C)^2 \cdot T(C^t + \Delta C|C^t) = \frac{C^t(1-C^t)}{N} \tag{17}$$

A.2 Stochastic Differential Equation

The continuous-time limit gives the Itô SDE:

$$dC^t = sC^t(1-C^t)dt + \sqrt{\frac{C^t(1-C^t)}{N}}dW_t$$

where W_t is a standard Wiener process. Setting $s = \Delta(u)$ from the main text:

$$dC^t = \Delta(u)C^t(1-C^t)dt + \sqrt{\frac{C^t(1-C^t)}{N}}dW_t$$

A.3 Implications

The stochastic term $\sqrt{C^t(1-C^t)/N}$ drives:

- **Drift-dominated fixation**: Neutral variants ($\Delta = 0$) can fix by chance
- **Latency phases**: Before deterministic growth dominates, constructions may persist at low frequency due to stochastic fluctuations
- **Small-population effects**: In communities with small N, stochastic effects can overwhelm weak selection

The mean-field approximation in the main text emerges by taking $N \to \infty$, where the diffusion term vanishes and we recover the deterministic logistic equation. For finite populations, the full stochastic dynamics must be considered, particularly near $C^t = 0$ or $C^t = 1$ where drift effects are strongest.

B Formal Derivations

B.1 Measurement Model for Community Entrenchment

The latent variable $C^t(u)$ is estimated through confirmatory factor analysis:

$$\begin{pmatrix} A(u) \\ \log E(u) \\ P(u) \\ S(u) \end{pmatrix} = \begin{pmatrix} \lambda_A \\ \lambda_E \\ \lambda_P \\ \lambda_S \end{pmatrix} C^t(u) + \begin{pmatrix} \epsilon_A \\ \epsilon_E \\ \epsilon_P \\ \epsilon_S \end{pmatrix}$$

where factor loadings λ_i and error terms ϵ_i are estimated from data. Identifiability is achieved by fixing $\lambda_A = 1$ and constraining all loadings to be positive.

B.2 Smooth Noise-Penalty Function

The noise-penalty function detects systematic errors using a smooth sigmoid:

$$\text{NoisePen}(u) = \beta_{\max} \left[1 + e^{-k(\text{Surprisal}(u) - \theta)}\right]^{-1}$$

where:

- $\beta_{\max}$ – maximum penalty for high-surprisal forms
- k – steepness of the transition (larger = sharper)
- θ – surprisal threshold where penalty = $\beta_{\max}/2$

This continuous function ensures smooth likelihood surfaces for parameter estimation while capturing the intuition that highly surprising forms with systematic repair patterns resist entrenchment.

B.3 Bifurcation Analysis

Actuation occurs when $\Delta(u) = 0$. Near this critical point:

$$C^t \approx C^* + Ae^{\lambda t}$$

where $\lambda = \Delta'(u^*) \cdot C^*(1 - C^*)$ determines growth rate post-bifurcation.

B.4 Processing Cost Function

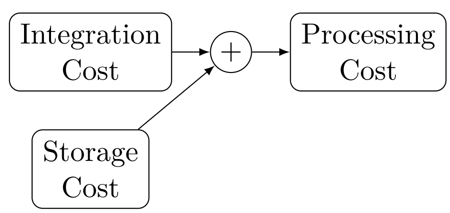

Figure 4: Computation of the processing-cost term.

Following Gibson (2000), Dependency Locality Theory:

$$\text{PCost}(u) = \sum_i \text{IC}(w_i)$$

where the integration cost at each word is:

$$\text{IC}(w_i) = \sum_{d \in \text{deps}(w_i)} |d|$$

with $|d|$ counting intervening discourse referents.

C Turkish Harmony Case Study

Turkish distinguishes lexical disharmony (tolerated) from allomorphic harmony violations (ungrammatical).

C.1 Stem-Internal: Phonology Only

(18) *doktor* 'doctor' (disharmonic stem, grammatical)

No morphosyntactic feature unrealized, so $G(u) = 1$ despite phonological markedness.

C.2 Suffixal: Morphosyntactic Requirement

(19) a. *kitap-lar* 'books' (harmonic, grammatical)
b. * *kitap-ler* (intended 'books', harmony violation)

The suffix must copy [±back] from stem. Wrong vowel leaves feature unrealized:

- Mapping succeeds (plural selected)
- $K(u) = 0$ (feature not realized)
- $C^t(u) = 1$ (everyone knows plural)

Therefore $G(u) = 0$, categorically ungrammatical.

C.3 Theoretical Implications

Turkish harmony shows morphosyntax-phonology interface effects on grammaticality. Only violations preventing feature realization trigger ungrammaticality; pure phonological dispreference affects only subjective feeling $F(u)$.

References

Baxter, Gareth J. et al. 2006. Utterance selection model of language change. *Physical Review E* 73(4). 046118. DOI: 10.1103/PhysRevE.73.046118.

Bergs, Alexander. 2025. From constructional innovation to linguistic change. *Functions of Language* 32(1). 16–42. DOI: 10.1075/fol.23013.ber.

Blythe, Richard A. & William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 88(2). 269–304. DOI: 10.1353/lan.2012.0027.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Futrell, Richard, Roger P. Levy & Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language* 96(2). 371–412. DOI: 10.1353/lan.2020.0024.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita & Wayne O'Neil (eds.), *Image, language, brain*, 95–126. Cambridge, MA: MIT Press.

Gibson, Edward, Leon Bergen & Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences* 110(20). 8051–8056. DOI: 10.1073/pnas.1216438110.

Giora, Rachel. 1997. Understanding figurative and literal language: the graded salience hypothesis. *Cognitive Linguistics* 8(3). 183–206. DOI: 10.1515/cogl.1997.8.3.183.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

Hofmeister, Philip, Laura Staum Casasanto & Ivan A. Sag. 2014. Processing effects in linguistic judgment data: (super-)additivity and reading span scores. *Language and Cognition* 6(1). 111–145. DOI: 10.1017/langcog.2013.7.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lai, Eric YongMing et al. 2013. The OPT-ional phenomenon in Singapore English: A corpus-based approach using time annotated corpora. *Procedia-Social and Behavioral Sciences* 95. 5th International Conference on Corpus Linguistics (CILC2013), 431–441. DOI: 10.1016/j.sbspro.2013.10.666.

Nowak, Martin A., Natalia L. Komarova & Partha Niyogi. 2001. Evolution of universal grammar. *Science* 291(5501). 114–118. DOI: 10.1126/science.291.5501.114.

Pearl, Judea. 2009. *Causality: models, reasoning, and inference*. 2nd edn. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511803161.

Reynolds, Brett. 2025a. Grammaticality as kind: ontology, epistemology, and empirical pay-off. manuscript in preparation.

Reynolds, Brett. 2025b. Habitus–conditioned grammaticality: A Bourdieusian extension of mmmg. Manuscript in preparation.

Reynolds, Brett. 2025c. Independent relative *whose*: From syntactic gap to pragmatic constraint. Under review.

Schütze, Carson T. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Berlin: Language Science Press. DOI: 10.17169/langsci.b89.100.

Snyder, William. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry* 31(3). 575–582. DOI: 10.1162/002438900554479.

Verhulst, Pierre François. 1838. Notice sur la loi que la population suit dans son accroissement. *Correspondance Mathématique et Physique* 10. Reprinted in *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 18, 1–42, 1845, 113–121.

Yang, Charles. 2000. Internal and external forces in language change. *Language Variation and Change* 12(3). 231–250. DOI: 10.1017/S0954394500123014.