De-idealizing the asterisk: Grammaticality as conditioned stability

Brett Reynolds
Humber Polytechnic & University of Toronto
brett.reynolds@humber.ca

Abstract

The asterisk has done foundational work in theoretical linguistics, but it also hides a persistent conflation. The same diacritic is used (i) to mark strings that defeat structural analysis, (ii) to mark structurally viable utterances whose values don't stabilize under the constraints of an interpreted situation, (iii) to mark interpretable forms that still aren't in the community's repertoire, and (iv) to mark cases that are grammatically well-formed but feel unacceptable for reasons of processing or ideology. One sign, four diagnoses: crash, clash, exclusion, and unease. This article argues that the resulting competence–performance–usage triangulation hasn't resolved the impasse because grammaticality has been asked to answer several distinct questions at once.

Moving forward, I propose a minimal state theory that reconceptualizes grammaticality as conditioned stability of form–value relations within communicative situations. Grammatical status depends on (i) mapping viability, (ii) interpretive coherence, and (iii) repertoire status. Distinguishing grammatical status from the feeling of ungrammaticality makes principled sense of classic dissociations between acceptability ratings and repertoire membership. The proposal yields operational diagnostics for separating coherence failures from repertoire exclusion, motivates an opportunity-normalized notion of negative evidence, and states concrete conditions under which the framework would be disconfirmed.

Keywords: grammaticality; acceptability; form–value relations; norms; preemption; processing; communicative situation

1 Introduction

Every competent speaker of English can tell that **Can the have running* is impossible. (So is **bnick* – but that failure is phonotactic, not syntactic; the two cases share the property of defeating analysis.) What is that impossibility a judgement of: structure, value, repertoire, or discomfort? The asterisk answers as if these were one. True structural crashes are rare – most asterisked examples in syntax are structurally analysable – but even this small class needs to be distinguished from the others. The cluster in (1) suggests why.

(1) a. **Can the have running?*
b. *Colorless green ideas sleep furiously.* (Chomsky, 1957)

c. **I've finished it yesterday.*
d. *?I saw Joan, a friend of whose was visiting.*
e. *The bread the baker the apprentice helped made is delicious.*
f. A: *How old are you?* B: **I have 25 years.*
g. **Which did you buy car?*

These items share the folk verdict that "something's wrong", but the wrong isn't the same. Example (1a) defeats structural analysis,[1] while (1b) is structurally impeccable but conceptually bizarre, even if a construal can be recovered. (1c) is interpretively transparent but clashes in temporal value: tense and modifier pull in different directions. For many speakers, (1d) is not confidently rejected so much as held at arm's length – judged marginal, uncertain, or simply unfamiliar in the repertoire. (1e) is often rejected in spontaneous use but becomes acceptable once a parse is stabilized, suggesting a processing-driven illusion. (1f) is viable and interpretable but isn't in the repertoire of the relevant English norm-centres, despite being ordinary in French and Spanish. (1g) is short and interpretable but treated as categorically excluded. Same verdict, different failure modes.

In short, the asterisk has been doing at least four jobs, and it has been doing them under a single typographic hood: it marks structural crash (1a); it marks interpretive incoherence (1c); it marks repertoire exclusion (1f), (1g); and it marks the feeling of anomaly driven by processing or ideology (1e). The asterisk is the dashboard light, not the mechanic. One glyph, multiple logics. That heterogeneity is the phenomenon.

Of course, linguistic practice already deploys multiple diacritics – *, ?, #, % – but their semantics varies across authors and traditions. Some use # for semantic anomaly, others for pragmatic infelicity; ? can mark gradient unacceptability or mere unfamiliarity. The taxonomy proposed here isn't a claim that linguists have ignored these distinctions; it's an attempt to give the existing diacritical toolkit a principled mapping onto underlying failure modes. The goal is to make explicit what the marks are supposed to be tracking, so that disagreements can be diagnosed rather than merely registered.

The history of grammaticality theory can be read as a sequence of attempts to compress such heterogeneity into a single explanatory core. Formal approaches treated grammaticality as categorical well-formedness; processing accounts treated gradience as performance; usage-based theories treated acceptability as the shadow of frequency and entrenchment; sociolinguistics treated grammaticality as norm-relative; experimental syntax refined measurement but didn't settle what's being measured. The result is a familiar triangulation in which the same data is alternately explained away as "competence", "performance", or "usage"; the triangle is stable, but the target keeps moving, often with little agreement on what would count as decisive evidence (Schütze, 2016; Sprouse et al., 2013).

This article is a contribution to the *Journal of Linguistics* section "Looking Back, Moving Forward". Looking back, I argue that the impasse persists because grammaticality has been asked to do the work of multiple distinct questions at once. Moving forward, I propose a minimal state theory that separates those questions—not to proliferate labels, but to make disagreement diagnos-

[1] Here "defeats" is meant in the restricted, behavioural sense made explicit in §4.2: no well-typed analysis is available to ordinary comprehension mechanisms even under prosodic scaffolding, explicit bracketing, and metalinguistic coaching that supplies intended constituency. "Rescuing" the string by positing unlicensed ellipsis or by stipulating missing lexical material changes the utterance type u; the relevant claim is that u itself lacks a viable mapping in the relevant conditioning state c.

able. The core proposal is that grammaticality is CONDITIONED STABILITY of form–value relations[2] within a communicative situation: grammatical status depends on (i) mapping viability (whether an expression-shape admits a well-typed structural analysis), (ii) interpretive coherence (whether the values encoded stabilize under the constraints live in the situation), and (iii) repertoire status (whether the form–value relation – especially at the operator stratum – is treated as a legitimate option in the relevant norm-centre). Parse, construe, belong: the asterisk has been made to answer all three. The same decomposition also clarifies how the FEELING OF UNGRAMMATICALITY arises as a metacognitive signal whose sources include, but aren't exhausted by, grammatical status; this distinction explains why some constructions feel ungrammatical while being licit, and why some illicit constructions escape detection (Fanselow, 2021).

The structure is as follows. Section 2 diagnoses the impasse by reviewing what the asterisk has been made to mean. Section 3 isolates the multiple questions that have been collapsed into one label. Section 4 introduces the state theory: conditioning states, the three constitutive quantities, and the stability score. Section 5 works through diagnostic profiles that the model predicts. Section 6 addresses evidence and measurement, including a worked opportunity proxy. Section 7 frames key questions for future research. Section 8 states what would count against the framework. Readers who want the core model quickly can focus on §§4.2–4.3, Table 3, and §8.

2 Looking back: what the asterisk has been made to mean

The modern theoretical role of grammaticality was shaped by the mid-century identification of grammar with a formal system generating a set of well-formed expressions. In this tradition, grammaticality is a categorical membership fact: a string is grammatical iff it's generated by the grammar (Chomsky, 1957). This view captures the hard edge of cases like (1a), where the system crashes before any stable analysis is available. It also provides a clean division of labour: semantics and pragmatics interpret outputs; performance systems realize them.

The cost of this idealization is that it forces the field to treat gradience as epiphenomenal. The competence–performance distinction (Chomsky, 1965) allowed formal theory to preserve categorical grammar by relocating variability to processing and attention, but the move is methodologically hazardous: once invoked, it can immunize the grammar from counterevidence by labelling inconvenient data as performance noise (Schütze, 2016, p. 71). Much subsequent work can be read as a search for principled ways to reintroduce gradience without abandoning the insight that some failures are genuinely categorical.

Meaning, coherence, and the limits of well-formedness present a second theme. Chomsky's (1b) was designed to show that structural well-formedness doesn't reduce to semantic plausibility. That point remains foundational: a theory that equates grammaticality with "interpretability" will misclassify many robust structural constraints. But (1b) also revealed a complementary fact: humans routinely accept structurally well-formed utterances whose values are conceptually odd, while rejecting other utterances whose intended interpretation is transparent. This tension motivated a long tradition of work linking acceptability to interpretive pressures, including semantic motivation for constraints (Lakoff, 1971; McCawley, 1968) and constructional meaning (Goldberg, 1995).

[2] I use VALUE for what a form conventionally contributes – primarily meaning, but extending to phonological and distributional regularities. Value is relational and contrastive: defined by opposition within a system, not by speaker intention (Saussure, 1916).

These traditions didn't establish that meaning replaces grammar; rather, they showed that the stability of interpretation is itself a locus of constraint. An utterance may be structurally viable but fail because the values encoded by its parts can't be reconciled under the constraints that are live in a situation. The present perfect plus *yesterday* in (1c) is a canonical case (Huddleston & Pullum, 2002, pp. 140–141): the intended meaning is obvious, but the morphosyntactic temporal value conflicts with the adjunct anchoring. Conversely, in English at least, many lexical clashes are tolerated as long as they don't implicate morphosyntactic value.

Processing and the reallocation of gradience constitute a third response. If grammar is categorical but judgements are gradient, one obvious move is to treat gradience as a function of processing. The processing literature has supplied a large inventory of robust effects – dependency locality, interference, garden-path reanalysis – that depress ratings and slow reading times for structures that are otherwise analyzable (E. Gibson, 2000; Grodner & Gibson, 2005). Classic centre-embedding examples like (1e) are often treated as the poster children: they are grammatical in the sense of analyzable and interpretable, but they trigger strong negative responses because incremental parsing is strained.

Processing explanations, though, don't exhaust the landscape. Certain constructions remain sharply rejected even when short and interpretable, and even when repeated exposure doesn't improve ratings. The literature on satiation and adaptation was partly motivated by precisely this need (Reynolds, 2025; Snyder, 2000, 2022): some degraded structures improve with exposure, others don't, and the difference can't be reduced to length or memory load alone. While processing accounts for why such structures feel ungrammatical, it doesn't, by itself, constitute a theory of grammatical status.

Usage, norms, and the social life of grammaticality provide a fourth strand. Usage-based approaches shifted attention to the role of frequency and entrenchment: speakers learn the distributions of forms, and those distributions shape what feels acceptable (Bybee, 2006, 2010; Reynolds, 2026c). A key advance in this tradition is the recognition of PREEMPTION: a form can be rejected because a competitor is consistently selected in the same niche, even if the discarded form remains structurally possible (Goldberg, 2011). The contrast between *I'm 25 years old* and **I have 25 years* in English illustrates the point: the latter is transparent and structurally viable, but is systematically excluded from the repertoire of the relevant norm-centres.

Sociolinguistic accounts, meanwhile, emphasize that grammaticality resides not in the abstract properties of a language, but in a community's normed repertoire (Labov, 1972). Indexical values attached to forms can shift what a situation admits to the repertoire, and speakers routinely disagree about what counts as "the" grammar because they construe different norm-centres as relevant (Eckert, 2012; Silverstein, 1976). Far from an embarrassment, this constitutes part of the phenomenon. The problem's that, in much theoretical practice, norm-relativity's treated as a complication external to grammar rather than as a constitutive feature of what grammatical status amounts to.

Probabilistic and gradient-competence models offer a fifth response by rejecting the categorical premise outright. Stochastic grammars assign probabilities to derivations rather than set membership; variable-rule frameworks model inherently gradient constraints; recent probabilistic syntax treats grammaticality as a continuous variable shaped by the learner's inductive generalizations (Bresnan et al., 2007; Manning, 2003). These approaches don't relocate gradience to performance – they build it into the grammar itself. The present proposal shares their empirical seriousness about gradience while preserving a distinction between objective status and subjective ratings.

Constraint-based architectures – Optimality Theory and its weighted variants – occupy related

terrain (Legendre et al., 2001). In classical OT, categoricality emerges from strict ranking; in Harmonic Grammar and MaxEnt models, weighted constraints yield gradient well-formedness as a function of constraint violation profiles. The key point for present purposes is that these frameworks treat gradience as a property of the grammar's output rather than as measurement noise. The decomposition proposed below is compatible with constraint-based internals while adding explicit conditioning on community and situation.

Finally, experimental syntax has increasingly distinguished multiple judgement types without waiting for theoretical consensus (Schütze, 2016; Sprouse et al., 2013). Magnitude estimation, forced choice, and timed tasks yield different profiles for the same items, suggesting that "acceptability" isn't a unitary quantity. Recent work has argued that at least some of this variation reflects distinct cognitive sources rather than mere task noise (Featherston, 2005). The state theory proposed here aims to provide a principled home for such distinctions: if different tasks tap different components of the stability score, divergent profiles are expected rather than anomalous.

What each tradition captures, and what it leaves unresolved, maps onto the decomposition that follows: probabilistic models take gradience seriously but typically don't separate objective status from subjective ratings; constraint-based architectures locate gradience in the grammar but don't condition on community norms; experimental syntax has the measurement sophistication but needs a theory of what's being measured. The state theory proposed below tries to combine these strengths: gradience in the stability score, conditioning on community and situation, and distinct measurement channels for distinct components.

3 The impasse diagnosed: three questions collapsed into one label

Rather than noise, the heterogeneity in (1) is structural. The asterisk collapses four things, but only three are constitutive of grammatical status itself; the fourth – the feeling of ungrammaticality – is a distinct phenomenon that needs to be separated. The three constitutive questions are easy to state and easy to confuse: can the system map the form, can the values cohere, and can the community treat the result as a legitimate resource? Grammaticality theory has repeatedly attempted to treat grammatical status as a unified phenomenon when it is, in fact, the intersection of these three questions.

First, structural viability. Some inputs fail because no structural analysis is available that yields a well-typed morphosyntactic representation. In such cases, the failure is categorical and doesn't depend on meaning, social norm-centres, or processing effort; the analysis crashes. Example (1a) is emblematic: the category sequence prevents the construction of a viable constituent structure.

Treating this failure mode as real is non-negotiable: without it, the notion of grammar loses its basic explanatory purchase. The mistake lies in elevating this single prerequisite into the definition of grammaticality itself.

Second, interpretive coherence. Many strings are structurally viable but unstable in value. Sometimes the instability is semantic (temporal alignment, argument structure satisfaction); sometimes pragmatic or information-structural (topic/focus fit); sometimes indexical (social meaning clashes with footing). The common thread lies in the stability of a dominant construal under the constraints that are live in the relevant situation, rather than in a folk notion of "meaningfulness".

Example (1c) illustrates: the intended interpretation is obvious, but the morphosyntactic value encoded by the present perfect conflicts with the temporal anchoring provided by *yesterday*. The result is interpretive instability grounded in conventional form–value relations, rather than structural

nonsense.

Third, repertoire status. A third class of cases are structurally viable and interpretively coherent, but rejected because they aren't in the community's repertoire. Here the role of usage and norms is constitutive: the community hasn't conventionalized the relevant form–value relation as a legitimate option under the norm-centres that define the communicative situation.

Example (1f) is again emblematic. The form isn't nonsensical, and it's interpretable. Its rejection is a fact about English community conventions, not about universal cognitive limits. The same form is licit in other languages, demonstrating that the relevant factor is repertoire status, not viability or coherence.

A fourth label deserves separation: the feeling of ungrammaticality. The three components above are constitutive for grammatical status. But speakers' judgements also reflect a FEELING OF UNGRAMMATICALITY: a metacognitive negative signal triggered by instability or high repair cost. This feeling is an important object of study, but it isn't identical to grammaticality. It yields false positives, where licit constructions feel bad, and false negatives, where illicit constructions pass undetected (Fanselow, 2021). Equating ratings with grammatical status invites conceptual confusion.

Four concepts need to be kept apart. APPROPRIATENESS is the genus: the fit between a form and the context in which it's used. GRAMMATICALITY is one species of appropriateness – the coupling between grammatical form and the values it conventionally expresses. It has a fact of the matter – but the fact in question is about community coordination, not Platonic membership. Either the form–value relation is stably treated as a legitimate resource by the norm-centred population for the relevant conditioning state, or it isn't. This is a claim about collective practice, not about abstract linguistic objects; it's realist in the sense that the coordination state is determinate even when our measurements are noisy, not in the sense that grammaticality floats free of speakers.[3] ACCEPTABILITY is the measurement channel: how speakers rate utterances, informed by grammatical status but also by processing factors, repair costs, and ideological filtering. CORRECTNESS is the prescriptive overlay: what gatekeepers enforce, what gets codified and moralized – often an ideologized version of one variety's appropriateness norms imposed as if universal. Fit, status, report, enforcement: related, but not the same. A predictable objection to the framework below is that C_t (repertoire status) merely relabels prescriptive correctness. The answer is no: correctness concerns what should be enforced; repertoire status is a constitutive fact about what a norm-centred population actually treats as a legitimate resource. Enforcement can distort evidence, but the repertoire state isn't the prescription (Pullum, 2019, for parallel distinctions).

Three cases populate the logical space. First, *Me and him went to the store*: in-repertoire for many speakers in informal registers (C_t high in those norm-centres), but proscribed by gatekeepers as "incorrect" – high repertoire status, low "correctness". Second, **Whom did you see him?*: a resumptive object after fronted *whom* – both proscribed and genuinely out-of-repertoire for most varieties (C_t low), so prescription and repertoire status align. Third, *I ain't got none*: fully in-repertoire in many vernacular varieties (C_t high within those norm-centres), but likely to trigger negative affect for ideological reasons in formal settings – high repertoire status, low ratings under gatekeeping framings. The triad shows that C_t isn't correctness: the same form can be in-repertoire but proscribed, out-of-

[3] The distinction between ontological fact and epistemic access is developed in Reynolds (2025, ch. 5), which argues that category boundaries are structurally determinate but located at thresholds we cannot finitely specify; gradient judgments arise from discrete categories filtered through processing noise, not from gradient membership.

repertoire and proscribed, or in-repertoire but ideologically disfavoured.

Ideology enters by two pathways, not one. First, it can be constitutive of C_t: sustained policing can reshape which form–value relations a community treats as legitimate, so that ideology partly determines repertoire status over time. Second, it can contaminate ratings: ideological filtering can depress acceptability for forms that are in fact in the repertoire, producing false negatives in the measurement channel. The architecture has to reflect both pathways. The first is etiological (ideology shapes C_t trajectories); the second is observational (ideology biases the ungrammaticality signal).

The remainder of the paper proposes a minimal state theory that makes these distinctions explicit, thereby clarifying what it is for an utterance type to be grammatical *in a communicative situation*.

4 Moving forward: grammaticality as conditioned stability of form–value relations

The state theory commits to just three things: analyzability, stability of construal, and repertoire status – each relativized to a construed situation. The notation that follows is bookkeeping, not a new philosophical burden: it's there to make the distinctions explicit.

4.1 Conditioning states and communicative situations

Consider a speaker who says *I seen it*. In a classroom presentation, this is likely to be heard as ungrammatical; at lunch with friends, it may pass without comment. The string hasn't changed; what's changed is which norm-centre is in play and what's at stake. This is what conditioning captures.

Let c be a conditioning state: a construed communicative situation together with whatever norm-centre is treated as relevant (Wiese, 2023). The point isn't to reify c as a fixed external context; interlocutors can misalign about which c is in force, and c can be renegotiated. In an experimental setting, c is partly latent: participants infer which conditioning state is in play from framing cues, and observed responses are expectations over participant-specific inferred c's. The modelling commitment is simply that grammatical status is always assessed relative to some such conditioning.

Rather than an optional sociolinguistic add-on, this move is the minimal way to state the empirical fact that grammars are socially situated repertoires: the same speaker can treat different resources as in-repertoire in different situations, and different speakers can rationally disagree about repertoire membership when they construe different norm-centres. This is the core of the realist commitment: grammaticality isn't an abstract property of the string, but a measurable state of the relation between form, value, and agents in a constructed situation.

The conditioning state c can be decomposed into at least three anchors (Reynolds, 2026a):

- Situation (S): the here-and-now interactional frame – activity type, medium, footing, institutional context.
- Ascription (A): what the speaker is treated as – the social categories assigned by self and others, which condition expectations about baseline repertoire.
- Identification (I): whose norms are being oriented to – the reference population the speaker treats as the standard for what counts as legitimate.

Together these yield a conditioning vector $c \approx \langle S, A, I \rangle$. Situational stakes are modelled separately as part of the decision regime: holding c fixed, stakes tune the criterion τ (and therefore the categorical verdict G_t) rather than directly conditioning map, K, or C_t. For compactness I continue to

write $\tau(c)$, but where stakes is manipulated independently the intended dependence is $\tau(c, \text{stakes})$. The decomposition matters because apparent disagreement about grammaticality often reflects misalignment in A or I rather than genuine conflict about the state of the form–value relation. The same token can be processed as dialectal (in-repertoire under one ascription) or as an error (out-of-repertoire under a different norm-centre), depending on which conditioning anchors the listener infers. This is why the classroom/lunch contrast for *I seen it* isn't just about formality (S); it's also about whose norms are in play (I) and what categorization the listener assigns to the speaker (A).

4.2 Three constitutive quantities

For an utterance type u in conditioning state c at time t, define three state quantities.

The first quantity is mapping viability. Let $\mathsf{map}(u, c) \in \{0, 1\}$ be a binary indicator of whether there exists at least one viable morphosyntactic analysis for u in c for which there's a well-typed representation (where 1 is viable and 0 isn't). map is intended to capture genuine analyzability failure and only that. It's the categorical prerequisite highlighted by the well-formedness tradition – the "entry ticket" to the system – but it doesn't on its own guarantee either interpretive coherence or repertoire membership. Many ungrammatical strings remain easily parsed and "interpreted" in a folk sense.

True $\mathsf{map} = 0$ cases are rare. Most strings that linguists mark with an asterisk are structurally analyzable; their failure lies in coherence or repertoire, not in parsing. The evidence for $\mathsf{map} = 0$ is behavioural: persistent comprehension failure even under supportive conditions – prosodic scaffolding, explicit bracketing, unlimited time, and metalinguistic coaching that supplies intended constituency. If a string can be made to yield a stable parse by restoring normal processing conditions, then $\mathsf{map} = 1$ and the difficulty belongs to processing costs rather than to analyzability itself.

A predictable objection: any string can be "rescued" given enough ingenuity – posited ellipsis, coercion, metalinguistic stipulation. The response is that such rescues change the utterance type. The definition of u is intentionally conservative: u is the string as presented, with whatever prosody, constituency, and lexical identity the speaker supplies. If a hearer has to posit missing material, reparse as quotation, or invoke a repair frame to achieve analysis, the resulting object is a different u', and the claim is that the original u lacks a viable mapping. This is why the behavioural criterion matters: what counts is whether ordinary comprehension mechanisms yield a well-typed analysis for the presented string, not whether a theorist can manufacture one. The lead example (**Can the have running*) may seem rescuable via NP ellipsis or by treating *the* as a deictic nominal; the point is that such moves don't describe how English speakers actually process the string – they describe analyst repairs that change the input.

The second quantity is interpretive coherence. Let $K(u, c) \in [0, 1]$[4] represent the stability of interpretation: the degree to which the utterance yields a dominant, non-contradictory construal under the constraints live in c (ranging from 0, complete instability, to 1, perfect coherence). Formally, K can be modelled as concentration of a distribution over candidate construals; for present purposes, the important point is that K is distinct from map. Structural viability doesn't guarantee coherence.

One concrete operationalisation separates construal multiplicity from construal stability. Present a target utterance in a supporting scenario, elicit a best-interpretation choice (or a paraphrase that is then clustered into a small set of readings), and then probe coherence *conditional on*

[4] The curly braces $\{0, 1\}$ denote a two-member set (exactly 0 or 1); the square brackets $[0, 1]$ denote the continuous interval from 0 to 1 inclusive.

the selected construal using inference questions or consistency checks. Low K is diagnosed when participants who have selected the same construal still show high rates of contradiction, revision, or dispersion in downstream inferences; stable ambiguity, by contrast, may yield multiple construal clusters at the population level while remaining high-K within each cluster. Population-level entropy over construal choices therefore diagnoses ambiguity or norm-centre mixture, while within-construal instability diagnoses low coherence. To decouple coherence from legitimacy policing, paraphrase tasks should be run under a "treat as legitimate dialect resource" framing, so that K measures stability of interpretation rather than willingness to engage with a stigmatised form.

The third quantity is repertoire status. Let $C_t(u, c) \in [0, 1]$ track whether the form–value relation u is a stabilised coordinative resource in the norm-centred repertoire relevant for c: whether there's a stable population-level practice in which producers use u for its conventional job and consumers treat it as an ordinary option (rather than as a slip, transfer artefact, or alien code) under the conditions that normally obtain for that repertoire. As an epistemic model of our access to this state, we represent uncertainty about repertoire membership by a population-level posterior probability: the probability that an individual drawn from the relevant norm-centred population treats u as a legitimate resource in c (where 1 represents universal acceptance and 0 indicates total exclusion). Corpus counts and task responses are evidence about C_t; they aren't what C_t consists in. This quantity is related to what usage-based work calls entrenchment, but it's explicitly conditioned on c and includes normative dimensions that pure entrenchment doesn't capture.

C_t is where norms live. It's also where many apparently categorical exclusions can be located without positing hard representational bans: a form can be structurally viable and interpretable while being near-universally excluded from the repertoire in a given situation.

The $S/A/I$ decomposition of c introduced above clarifies what fixes "the community" for a given evaluation. Identification (I) is the natural anchor for whose repertoire counts in C_t: the reference population is whoever the speaker is orienting to. Situation (S) and stakes are the natural anchors for the decision regime $\tau(c)$: high-stakes institutional contexts raise the threshold. Ascription (A) explains a major source of apparent inconsistency: the same string can be treated as in-repertoire when attributed to one ascribed group and as an error when attributed to another, even if the underlying variety grammar is the same. That isn't a change in map or K; it's a change in how listeners map tokens to norm-centres.

Symbol	What it tracks	Diagnostic evidence
map	Structural analyzability (0/1)	Persistent parse failure under scaffolding; no stable category assignment
K	Interpretive coherence (0–1)	Paraphrase dispersion, construal instability
C_t	Repertoire status (0–1)	Production rates, "would you say this?", corpus frequency normalized by opportunity

Table 1: The three constitutive quantities at a glance.

4.3 A stability score and a membership predicate

Define a graded stability score:

$$\widetilde{G}_t(u,c) = \mathsf{map}(u,c) \cdot K(u,c) \cdot C_t(u,c) \in [0,1]. \tag{2}$$

This multiplicative scoring means that if any single component is zero – if the mapping fails, if interpretation is impossible, or if the form isn't in the community's repertoire – the entire relation is ungrammatical. Stability underwrites gradience: lowering any component reduces the overall score.

In plain terms: the product rule says that being in-repertoire can't make up for being incoherent, and being perfectly coherent can't make up for being out-of-repertoire. Deficits compound rather than average out.

This decomposition reflects a broader organization of linguistic infrastructure. Expression-shape constraints (phonotactics, morphotactics) regulate whether an utterance is recognizable as a token of the system; their violation yields "not a word". Operator-like constraints – closed-paradigm contrasts that configure public update, allocate participant roles, and authorize uptake – are targeted by K (for value coherence) and C_t (for community repertoire); their violation yields "you can't say that". Payload resources (open-class lexicon, indexical stance) remain negotiable and extensible; their misuse invites clarification or social judgment, not structural rejection. The stability score $\widetilde{G}_t$ integrates across these levels: a form that crashes at any level is unstable, but the *type* of instability differs diagnostically.

Communities also often treat grammaticality as a categorical membership fact: either a resource is in the repertoire or not. Model this by thresholding:

$$G_t(u,c) = \mathbb{I}\Big[\widetilde{G}_t(u,c) \geq \tau(c)\Big], \tag{3}$$

where $\tau(c)$ is a situation-specific decision criterion. The point is that $\tau(c)$ is a property of how strict the situation is about what counts as "in" the repertoire. High-stakes institutional contexts can set a high threshold; low-stakes in-group contexts can set a lower one. The graded score $\widetilde{G}_t$ is the primary grammatical-status object – the latent state. The categorical predicate G_t is a regime-dependent classification: how communities or experiments convert stability into a binary verdict. This keeps the realist commitment (the coordination state $\widetilde{G}_t$ is determinate even when our measurements are noisy) while acknowledging that the categorical label depends on who's drawing the line and what's at stake.

To illustrate: in a classroom presentation, using a stigmatized dialectal form risks being marked down; the threshold for "grammatical enough" rises. At lunch with friends, the same form may index solidarity; the threshold drops. Same form, same $\widetilde{G}$, different verdicts.

A concern is that $\tau(c)$ might immunize the theory if it can vary freely. Two constraints matter. First, $\tau(c)$ isn't construction-specific: it's fixed for a conditioning state and therefore shifts the boundary for *all* utterance types evaluated in that state. Adjusting τ to rescue a single problematic case entails collateral predictions for a broad set of anchor items. Second, $\tau(c)$ can be motivated by a standard decision-theoretic rationale in which classification losses differ by situation. Let $L_{\text{FA}}(c)$ be the loss of treating an item as in-repertoire when it is not, and $L_{\text{FR}}(c)$ the loss of treating an item as not-in-repertoire when it is. A natural constraint is

$$\tau(c) = \frac{L_{\text{FA}}(c)}{L_{\text{FA}}(c) + L_{\text{FR}}(c)}.$$

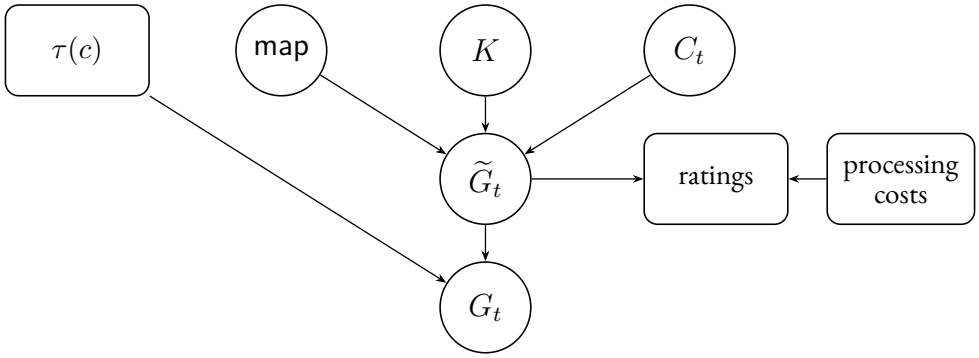

Figure 1: The minimal state architecture. Grammatical status $\widetilde{G}_t$ combines mapping, coherence, and repertoire status. Acceptability ratings reflect $\widetilde{G}_t$ filtered through processing costs, while categorical membership G_t involves a situation-specific threshold $\tau(c)$.

This ties τ to independently characterizable properties of the communicative situation (stakes, institutional norms, gatekeeping pressure). Empirically, $\tau(c)$ can be estimated by calibrating participants on an anchor set spanning clear in-repertoire and clear not-in-repertoire items for the target c, rather than being tuned post hoc to accommodate the construction under dispute.

This formalizes an intuition that's often stated informally but rarely built into the state theory: what counts as "grammatical" for practical purposes depends on the decision regime of the situation, not just the resource itself. Figure 1 summarizes the minimal architecture.

4.4 Why this combination rule isn't decorative

(If you accept the decomposition, the point of this subsection is just to show that the multiplication yields discriminable interaction predictions. Skim if pressed.)

The three-way decomposition is the theoretical commitment; the choice of a specific combination operator is a modelling decision. Substantively, the product approximates what successful coordination looks like when it depends on three independently failing gates: the form has to be recognisable (map), its values have to cohere (K), and the community has to treat the relation as available (C_t). If any gate fails, coordination fails. The operator should be further constrained by desiderata that make it empirically non-trivial.

First, the core is non-compensatory: mapping failure, catastrophic incoherence, or categorical repertoire exclusion should each be sufficient to drive grammatical status to zero in the relevant c. This excludes simple weighted sums as a model of grammatical status, since they permit a high value on one dimension to compensate for near-zero on another. Second, the graded score should reflect compounding instability: two moderate deficits should typically be worse than either deficit alone. Third, the operator should be monotone in each argument, and it should allow a transparent generalization to relative weighting if later work justifies it.

Several standard operations meet the non-compensatory constraint. The minimum operator,

$$\widetilde{G}_t^{\min}(u, c) = \min\{\mathsf{map}(u, c), K(u, c), C_t(u, c)\},$$

treats the weakest link as decisive. This makes a clear prediction: once one component is identified as the bottleneck, further degradation elsewhere shouldn't matter for the objective score. At the other

extreme, a weighted sum predicts systematic compensation:

$$\widetilde{G}_t^{\Sigma}(u,c) = w_{\mathsf{map}}\mathsf{map} + w_K K + w_C C_t,$$

which is often plausible as a model of subjective ratings but is a poor fit for a state theory of grammatical status precisely because it allows a community to "make up for" incoherence by repertoire status alone.

The product rule adopted in (2),

$$\widetilde{G}_t^{\times}(u,c) = \mathsf{map}(u,c) \cdot K(u,c) \cdot C_t(u,c),$$

is the simplest operator that's non-compensatory and compounding. It also has a useful interpretive property: in log-space, the components contribute additively ($\log \widetilde{G} = \log \mathsf{map} + \log K + \log C_t$), which aligns naturally with an evidence-accumulation picture in which distinct sources of instability contribute independent penalties. If future work motivates differential weighting, the product generalizes straightforwardly to a weighted geometric form, $\mathsf{map} \cdot K^{\alpha} \cdot C_t^{\beta}$, with $\alpha, \beta > 0$.

The choice among min and $\times$ is empirically discriminable. Consider a factorial manipulation that independently lowers coherence and repertoire status while holding mapping constant: for the same morphosyntactic frame, introduce a mild value-clash (lowering K) and, independently, present the construction under a norm-centre that treats it as non-native or marginal (lowering C_t).[5] The minimum rule predicts that once either K or C_t is the bottleneck, the second manipulation shouldn't further depress the objective score; the product rule predicts a systematic interaction (compounding), since the combined manipulation reduces stability more than either alone. This is a substantive prediction about the structure of the state space, not a restatement of the verbal story.

An information-theoretic perspective clarifies why this matters. The multiplicative structure has a natural interpretation in terms of how linguistic contrasts contribute to interpretation. Some form–value relations occupy small, closed paradigms but cause large downstream consequences: clause type constrains which responses are relevant; polarity flips entailment relations; case and agreement constrain role assignment. These relations function as control settings – protocol headers rather than payload content – carrying few bits in themselves but causing large entropy reduction in the space of licit interpretations (Cover & Thomas, 2006; Shannon, 1948). A wrong value doesn't merely produce a surprising concept combination; it disrupts the mapping from form to publicly recognizable update. This is why a short, interpretable utterance like (1c) can trigger categorical rejection: the violation targets infrastructure, not content. The product rule captures this asymmetry: degradation in control-like dimensions (map, K for operator-relevant constraints, C_t for high-opportunity paradigms) compounds rapidly, while payload-level infelicities remain negotiable.

Empirical upshot: $K \times C_t$ manipulations should compound under the product rule; if they don't, the operator is wrong even if the decomposition stands.

A concrete design sketch makes this testable. Take a 2×2 factorial with K (high vs degraded) crossed with C_t (in-repertoire vs marginal). The K manipulation introduces a mild temporal or aspectual clash (e.g., present perfect with a definite past adverb vs simple past with the same adverb);

[5] The repertoire-status manipulation can be implemented by norm-centre framing (ingroup/dialect/resource vs error) and by register framing (a shift in S, hence in c, which can in turn shift C_t). Stakes framing is predicted primarily to shift the decision criterion τ.

the C_t manipulation shifts norm-centre framing (in-group vs formal/institutional). Measure acceptability ratings and production probability. The minimum rule predicts a ceiling effect: once either K or C_t is the bottleneck, the second manipulation shouldn't further reduce ratings. The product rule predicts an interaction: the high-K, marginal-C_t cell and the degraded-K, in-repertoire cell should both show moderate degradation, but the degraded-K, marginal-C_t cell should show superadditive degradation beyond what either deficit alone produces. If ratings in the double-deficit cell match the single-deficit cells (no interaction), the minimum rule is favoured; if ratings in the double-deficit cell are reliably worse than the worst single-deficit cell, the product rule is favoured. This is a realizable experiment, not a thought experiment.

A clarification about scope: the product rule is a constitutive description of the stability score at a fixed time – a snapshot. The three components are plausibly entangled in their dynamics: low K can depress production, which reduces evidence and can lower C_t over time; processing costs (outside $\widetilde{G}_t$) can feed back into C_t by discouraging use. These causal couplings belong to the etiological module (how states arise), not to the combination rule (what the state is at evaluation time). The product can be the right constitutive operator even if the components are dynamically correlated.

4.5 Separating coherence from repertoire status: operational criteria

A recurring worry is that coherence failure $K \approx 0$ and repertoire exclusion $C_t \approx 0$ may collapse into one another, since both yield low stability. The separation requires distinct signatures, not just distinct labels.

Low K (coherence failure)	Low C_t (repertoire exclusion)
Construal unstable; speakers disagree on meaning	Construal stable; speakers agree on meaning
Paraphrase dispersion high	Paraphrase agreement high
Repair-heavy, effortful interpretation	Readily interpretable
"What does that even mean?"	"I know what you mean, but we don't say that"
Diagnostics: paraphrase tasks, construal variability, RT to inference questions	*Diagnostics:* production probability, "would you say this?", corpus frequency / opportunity

Table 2: Separating coherence from repertoire status: predicted contrasts.

These diagnostics cut across the tempting verbal contrast between "values can't be reconciled" and "the community doesn't accept the reconciliation". In practice, the decisive question is whether the source of degradation is interpretive dispersion or repertoire exclusion.

Two additional signatures don't reduce to paraphrase comparison. First, a TREAT-AS-CODE manipulation: explicitly instruct participants that the form is a legitimate resource of the target norm-centre, then test whether coherence stabilises. If coherence remains unstable under this framing, that's evidence for low K; if coherence stabilises but production and "would you say this?" responses remain near-zero, that supports low C_t. Second, repair profiles (§5): mismatches targeting operator-like dimensions are predicted to elicit open-class repair initiation and explicit rejection, while payload mismatches elicit stance negotiation. A form that triggers "what does that even mean?" is failing on K; a form that triggers "you can't say that" is failing on C_t.

This is why (1c) is a useful but non-trivial diagnostic. Many speakers can recover the intended meaning of *I've finished it yesterday* with little difficulty, which pushes it toward a low-C_t profile (repertoire exclusion of a specific tense–adverb pairing) rather than a pure low-K profile. On the other hand, if an experimental design reveals systematic competition between two construals (a present-perfect reading vs a coerced simple-past reading), then the same item will show low K by exhibiting dispersion in paraphrase and inference tasks even when participants are instructed to treat the form as a legitimate dialectal resource. The framework is falsifiable here: it predicts that the K-diagnosis and the C-diagnosis diverge in their measurement signatures.

The expected correlation and what counts as dissociation deserve explicit statement. Low C_t and low K will often co-occur: unfamiliar forms can induce interpretive uncertainty, and forms that resist stable construal may fail to enter the repertoire. But the correlation isn't identity. A genuine dissociation is diagnosed when the two signatures diverge under controlled manipulation. Specifically: under a treat-as-legitimate framing, low-C_t items should show stable downstream inferences once a construal is selected (participants agree on what it means and draw consistent conclusions), whereas low-K items should continue to show within-construal instability (participants who select the same reading still contradict themselves on inference probes). If both signatures pattern together across all manipulations, the separation is empirically empty; if they diverge, the decomposition is supported.

A non-transfer L1 contrast illustrates the distinction. Consider the stigmatized double-modal construction *I might could do that*, common in some Southern US varieties but categorically rejected by speakers of other varieties. For speakers who lack the construction, $C_t \approx 0$: it isn't in their repertoire. But K is high: the intended meaning (epistemic possibility of ability) is transparent, and speakers readily paraphrase it as *I might be able to do that*. Contrast this with a genuinely incoherent sequence like *The meeting starts before it begins*, which resists stable interpretation: the temporal predicates contradict each other, producing construal instability (K low) that persists even when participants are explicitly instructed to find a coherent reading. The prediction is that *might could* under a treat-as-legitimate framing will yield high paraphrase agreement and consistent inference patterns (high K), while *The meeting starts before it begins* will yield paraphrase dispersion and inference inconsistency (low K), even when participants try to accommodate it. Both may be rejected, but the source of rejection differs: repertoire exclusion vs coherence failure. This is a within-L1 test of the separation, independent of transfer.

4.6 Grammaticality versus the feeling of ungrammaticality

The state theory above defines grammatical status via $\widetilde{G}_t$ and $\tau(c)$. Speakers' ratings often track a different quantity: a subjective ungrammaticality signal driven by low stability, processing costs, and ideological overlays.

A useful way to characterize this signal is as inverse conditioning. If speakers condition production on S, A, and I, then listeners can infer those conditioning anchors from observed forms – Bayes' theorem running in reverse. The feeling of ungrammaticality is naturally tied to surprisal relative to the listener's inferred conditioning model: hearing a form that's low-probability under the c the listener thinks is in force triggers the signal. Processing costs and ideological overlays layer on top, but the core input to the detector is $-\log P(u \mid c)$. This framing has two methodological payoffs. First, it gives a principled bridge from ratings to a measurement channel: ratings are observations of a detector whose input includes surprisal plus processing costs, not direct observations of C_t.

Second, it makes the later discussion of language models less risky: LMs approximate something like $P(u \mid \text{context})$ for some training-conditioned mixture of c's, which is naturally closer to "detector input" than to "truth about G_t".

This distinction predicts systematic dissociations:

- Licit but degraded: $\mathsf{map} = 1$, K high, C_t high, but processing costs depress ratings (classic centre embedding).
- Illicit but unnoticed: $\mathsf{map} = 1$ and the intended meaning is salient, so the ungrammaticality signal is weak even when a relevant coherence constraint is violated (agreement attraction and other slips in complex structures; Wagers et al. 2009).

Equating acceptability ratings with grammatical status conflates a state claim with a measurement channel (Reynolds, 2025, ch. 5). The methodological consequence is that claims about G_t should be supported by converging indicators, with ratings treated as evidence primarily about the ungrammaticality signal and only indirectly about repertoire status.

5 Diagnostic profiles: what different failures look like

Beyond being definitional, the value of a state theory lies in the diagnostic profiles it predicts. The decomposition in (2) yields a compact typology of recurrent instability modes. The typology reflects that an utterance can be structurally well-mapped and easily "interpreted" in a folk sense while remaining ungrammatical due to coherence failure or repertoire exclusion.

Profile	Canonical signature
$\mathsf{map} = 0$	Persistent parse failure under scaffolding; categorical rejection; no amount of context stabilizes meaning (1a).
$\mathsf{map} = 1$, $K \approx 0$	Value incompatibility; intended meaning might be guessable, but conventional form–value constraints in c block stabilization (1c).
$\mathsf{map} = 1$, K high, $C_t \approx 0$	Repertoire exclusion; interpretable but treated as not in the repertoire; often cross-linguistically variable (1f, 1g).
$\mathsf{map} = 1$, K high, C_t low/uncertain	Rarity/indeterminacy; weak consensus; high variance across speakers (1d).
$\mathsf{map} = 1$, K high, C_t high, but high processing cost	Illusory ungrammaticality; improves with guidance; ratings track repair cost more than status (1e).

Table 3: Recurrent diagnostic profiles as regions of the state space.

To see the table at work, revisit the opening cluster. *Can the have running* (1a) is row 1: $\mathsf{map} = 0$. *I've finished it yesterday* (1c) is row 2: the string parses, but temporal values clash (K low). *I have 25 years* (1f) is row 3: fully interpretable, but English doesn't have it in repertoire ($C_t \approx 0$). *A friend*

of whose (1d) is row 4: the opportunity set is small, so speakers are uncertain rather than categorical (Reynolds, 2024). *The bread the baker...* (1e) is row 5: licit but processing-heavy, producing illusory ungrammaticality. The table isn't ornamental; it partitions the puzzle set.

Two contrasts are key for the future research agenda: stable repertoire exclusion versus rarity, and objective status versus felt ungrammaticality.

5.1 Stable repertoire exclusion versus rarity

A raw corpus absence is compatible with two very different states. A construction can be rare because the opportunity set is tiny, leaving speakers with little evidence either way; or it can be rare because, despite a large opportunity set, it's systematically preempted by competitors, driving repertoire status toward zero. The independent relative genitive in (1d) plausibly belongs to the first class for many speakers: the configuration that would make it useful is itself rare, so the absence of tokens doesn't straightforwardly imply categorical exclusion.

Left-branch extraction in (1g) behaves differently. The communicative niche is common, competitors are available (*Which car did you buy?*), and speakers show robust categorical rejection. This profile is analyzed as near-zero repertoire status in the relevant norm-centres, consistent with a preemption-based trajectory (Goldberg, 2011; Reynolds, 2026c). In this view, categoricality needn't be located in `map`: the intended analysis can be available and interpretation can be coherent once stipulated, while the community treats the relation as excluded from the repertoire.

5.2 Illusory ungrammaticality and misattribution

Processing-driven illusions illustrate why the feeling of ungrammaticality can't be equated with grammatical status. Centre embedding (1e)'s analyzable and interpretable, but incremental parsing strains working memory and dependency integration, triggering strong negative affect (E. A. F. Gibson, 2026). Similarly, garden-path items can feel nonsensical until reanalyzed:

(4) *The old man the boats.* (Ritchie & Thompson, 1984)

A first-pass parse yields nonsense; reanalysis yields a coherent, licit structure. In such cases, ratings track repair difficulty, not repertoire status (cf. E. A. F. Gibson, 2026, on acceptability as an introspective report of processing cost). Conversely, illicit structures can pass unnoticed when meaning is compelling, yielding false negatives (Pullum, 2009).

The repair system provides converging evidence. When repair does occur, mismatches targeting operator-like dimensions (Reynolds, 2026b) – tense errors, agreement failures, clause-type confusions – are predicted to elicit open-class repair initiation (*what?*, *who did it?*) and explicit rejection, because they disrupt the publicly accountable control settings on which uptake depends. Mismatches targeting payload or indexical dimensions are predicted to elicit stance negotiation and accommodation (*did you mean...?*, *why are you talking like that?*), because the utterance's update potential remains intact even when its content or social positioning is problematic. This asymmetry is independent of the feeling of ungrammaticality: a processing-heavy but licit structure may feel terrible without triggering the repair profile associated with genuine operator failure.

The state theory predicts such dissociations whenever the ungrammaticality signal pools multiple sources of difficulty.

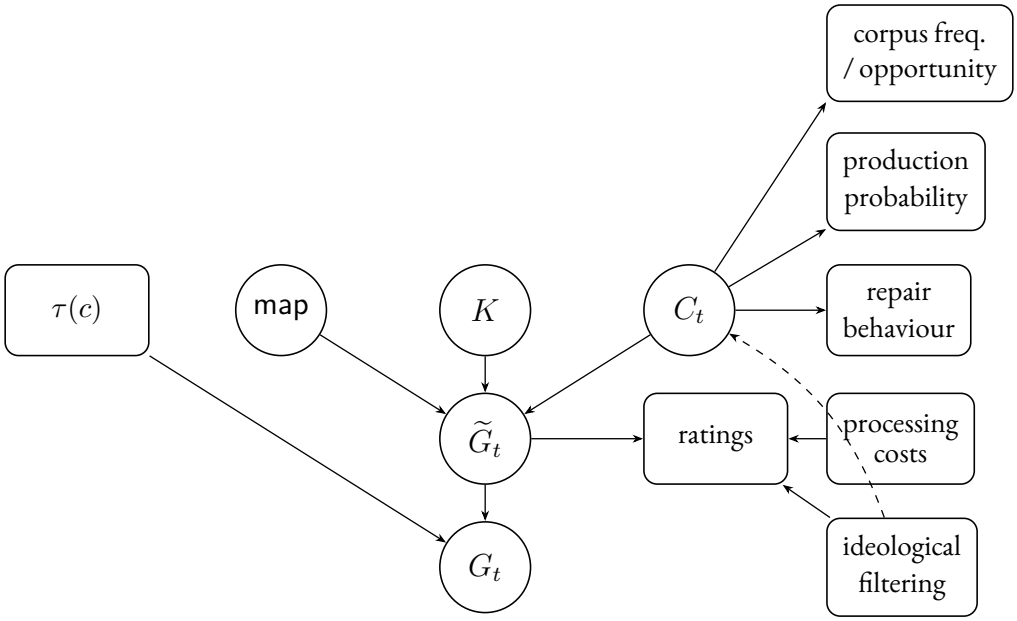

Figure 2: The medium-resolution architecture. Adds observable indicators for C_t and ideological filtering on ratings. Solid arrows represent synchronic generative influence; the dashed arrow indicates a diachronic etiological pathway (policing reshapes C_t trajectories over time). Inference runs opposite to arrow direction.

6 Evidence and measurement: what it would take to test the state theory

A "moving forward" programme has to specify what would count as evidence. The constitutive variables suggest a principled division of labour among data types (Figure 2).

For mapping viability, evidence comes from analyzability: whether speakers can assign a stable category structure, whether repairs consistently fail, and whether comprehension collapses even under supportive contexts. Structural crash cases are rare but diagnostically clean.

For coherence, evidence comes from interpretive stability under controlled manipulations of the relevant constraints (temporal alignment, argument structure, information structure, indexical consistency). Here experimental pragmatics and semantics supply tools for isolating which constraints are doing the work, while corpus work can reveal conventional distributional restrictions that track those constraints.

For repertoire status, C_t is latent and can't be inferred from ratings alone: not from ratings in isolation, not from a single corpus count, not from any one task. It has to be estimated from converging indicators: production probability in elicitation, corpus frequency normalized by opportunity sets, repair behaviour, recognition latency, and social evaluation under explicit norm-centre manipulations.[6] The state theory motivates an explicit measurement model for C_t in which acceptability ratings are treated primarily as observations of the ungrammaticality signal, not of repertoire status.

[6] A natural statistical commitment: model C_t via a hierarchical Beta–Binomial, treating production and acceptance responses as Bernoulli draws and entering opportunity-normalised corpus counts as exposure. Partial pooling across speakers and constructions accommodates the clustering structure. The point isn't to prescribe a single implementation but to commit to a recognisable class of estimator.

One central challenge involves operationalizing OPPORTUNITY. Preemption-based accounts require not only token counts but niche counts: how often the communicative job arises in the relevant c. A key empirical task for the moving-forward agenda is to develop operational definitions of niches for different constructions and to measure non-occurrence relative to those opportunities (for a corpus-based implementation using Bayesian partial pooling over construction-specific opportunity sets, see Reynolds, 2026c). The next subsection sketches the logic with a simpler case.

6.1 A WORKED OPPORTUNITY PROXY: AGE-STATING

Opportunity-normalisation is the hinge between mere corpus rarity and evidence of systematic exclusion. The general problem is that niches aren't directly annotated in corpora: we rarely observe "the speaker needed to express X" as an explicit variable. A workable starting point is to use competitor forms as a lower-bound proxy for opportunities. If speakers reliably realise a niche using an established competitor, then each observed competitor token witnesses an opportunity in which the target variant could, in principle, have been selected.

Consider (1f), **I have 25 years*, a construction with *have* taking age as object. The niche is: 'state a person's age in response to a direct or indirect enquiry'. Competitor realisations of this niche are easy to identify – all copular constructions with age as predicative complement: (i) *I'm 25*; (ii) *I'm 25 years old*; (iii) *I'm 25 years of age*. Each token of (i)–(iii) witnesses an opportunity: a speaker needed to state an age and chose one of these forms.

Let N^* be the competitor count in a corpus slice approximating the relevant c. Age-stating is common: a conversational corpus will contain thousands of tokens of (i)–(iii). If the *have*-construction were in the English repertoire – even as a rare option – we'd expect to find some tokens among those thousands. We find none, and that zero matters precisely because N^* is large. Compare a rarer niche: if the competitor count is 50, finding zero tokens of a target variant tells us almost nothing – the form might simply be uncommon. The evidential force of absence scales with opportunity.

The *have*-construction is entirely absent from L1 English data, despite being the productive pattern in French, Spanish, and other languages. The mapping is transparent (map and K pose no obstacle: hearers readily interpret *I have 25 years*), but $C_t \approx 0$ – the *have*-construction simply isn't part of the L1 English repertoire for this niche. Occasional tokens from L1 French speakers are predicted rather than problematic: C_t is conditioned on c, and the speaker's linguistic identification (I) is part of c. For the L1 English community, $C_t \approx 0$; for French-L1 speakers of English, it may be substantially higher. An L1 English hearer understands the utterance without difficulty – map and K are intact – but recognises it as outside the community repertoire. That recognition *is* the low-C_t judgment. The asterisk on (1f) is then not a brute intuition but a consequence of low C_t in a well-mapped, coherent construction – exactly the kind of diagnosis the model is designed to deliver.

A second illustration uses left-branch extraction (1g). The niche is: form a *which*-question targeting an NP object. Competitor realisations are abundant: *Which* $\langle N \rangle$ *did you* $\langle V \rangle$*?* and related wh-NP frames. Let N^*_{LBE} be the count of such competitor tokens in a corpus slice approximating c. If (1g) were merely rare, we'd expect at least occasional tokens once N^*_{LBE} is large. Instead, the target variant remains absent or vanishingly close to absent in the relevant norm-centres, despite massive opportunity and an easily available competitor (*Which car did you buy?*). This is the signature of a stable gap rather than low-opportunity rarity: high N^*, near-zero target rate. Corpus detectability matters here: the point is methodological (define a competitor set whose recall you can audit), not that any single corpus search is dispositive.

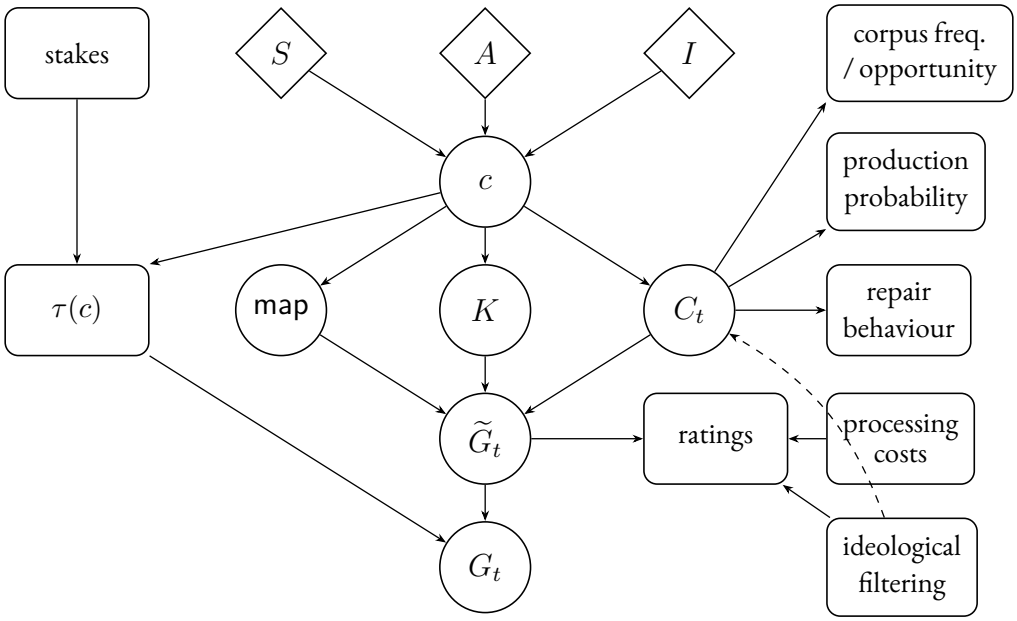

Figure 3: The full conditioned architecture. Shows how the conditioning state c is constituted by situation (S), ascription (A), and identification (I) anchors, and how it drives the state variables. Stakes drive the threshold $\tau(c)$ independently of c's content. The dashed arrow indicates ideology's etiological influence on C_t trajectories (cf. Figure 2).

This proxy operationalization is deliberately coarse, but it's already discriminating: it separates cases where "no tokens" is probative (large N^*) from cases where it isn't (small N^*). More generally, this is what a "moving forward" corpus programme has to provide: explicit definitions of competitor sets for niches and principled choices of corpus slices approximating c.

7 Key questions for future theoretical research

The state theory reframes several longstanding debates as tractable research questions.

How should conditioning states be operationalized? If grammaticality is conditioned, then specifying c isn't optional (Figure 3 shows the full conditioned architecture). Future work has to develop operational proxies for norm-centres and communicative situations: genre, medium, stance, audience design, institutional stakes, and community membership. An important prediction is that constructions whose status is driven by C_t will be more sensitive to c-manipulation than map-failures and many coherence-failures.

What are the right objects of repertoire membership? The theory treats u as an utterance type, but in practice the granularity of u matters. Is *gave the dog a bone* in-repertoire as a specific string, as an instance of the ditransitive construction, or as part of a broader caused-possession family? A moving-forward programme has to articulate principled criteria for individuating u in a way that makes the repertoire term empirically meaningful rather than vacuous.

What is the etiology of stable gaps? The present paper has remained mostly constitutive. The natural next step is an etiological module: a model of how $C_t(u, c)$ trajectories arise under positive evidence, error evidence, and opportunity-sensitive preemption. Instead of debating whether pree-

mption exists, the crucial question is its effective strength across niches and how it interacts with processing difficulty and social evaluation. This is where classic "categorical" constraints become a test case: the moving-forward claim is that at least some of them can be redescribed as stable repertoire exclusion sustained by strong preemption in robust opportunity sets. A companion paper developing this etiological module is in preparation.

A related question is which form–value relations attract sharp repertoire boundaries in the first place. The present framework is neutral on this, but a natural hypothesis is that repertoire policing clusters around OPERATOR contrasts (Reynolds, 2026b): closed-paradigm choices that configure public update, allocate participant roles, and constrain uptake – clause type, argument linking, tense–aspect where grammaticalized, evidential anchoring. If this is correct, then C_t trajectories are shaped not only by opportunity mass but by the functional load of the contrast: high-entropy-reduction dimensions attract categorical policing because a wrong value causes coordination failure even when the utterance is otherwise intelligible. This reframes the "categorical vs. gradient" debate as a question about which dimensions of the state space are operator-like, rather than about whether gradience is real.

A complementary etiological resource is coordination equilibria in the sense of evolutionary game theory (O'Connor, 2019). On this view, communicative situations are payoff structures, and repertoire boundaries stabilize because they solve recurring coordination problems. Some partitions become sharp and policed because category salience enables coordination: once a contrast is salient, speakers and listeners expect each other to respect it, and deviation is costly. This explains why C_t can remain near zero for forms that are structurally viable and interpretable – the coordination equilibrium excludes them. It also explains why certain boundaries resist erosion even under exposure: the equilibrium is self-sustaining because unilateral deviation is penalised. This game-theoretic module is compatible with the constitutive framework but adds an explanation of why some gaps are stable and others drift. The threshold $\tau(c)$ fits naturally into this picture: high-stakes situations are precisely those where coordination failure is costly, and institutions often encode the expected equilibrium as explicit gatekeeping.

How should typological generalizations be interpreted? If grammatical systems are normed repertoires shaped by stability dynamics, typological regularities are naturally viewed as recurring attractors in design space rather than as exceptionless laws. The task is to identify which combinations of form–value relations are robustly stable across lineages and which are contingent on local history and norm-centres. Large-scale typology becomes evidence about the global stability landscape rather than a direct route to categorical universals.

What role should language models play? Language models are now unavoidable instruments in linguistic practice. The state theory suggests a principled way to use them without mistaking their outputs for grammatical truth. If a model is treated as a proxy for the ungrammaticality signal, it may be useful for predicting processing difficulty and surprisal-like effects; if it's treated as evidence about repertoire status, it has to be grounded in opportunity-normalized distributions and norm-centre conditioning. The resulting agenda is methodological: what, exactly, are models approximating when they mimic human judgements, and which variable in (2) does that approximation correspond to?

8 What would count against this framework?

A framework that decomposes grammatical status into multiple components risks appearing too flexible unless each component is tied to independent evidence. The present proposal is disconfirmed, or at least seriously pressured, by any of the following patterns.

Each condition targets a specific earlier commitment: (1) targets the K/C_t separation (§4.5); (2) targets the opportunity methodology (§6.1); (3) targets conditioning and thresholds (§4.3); (4) targets the combination rule (§4.4); (5) targets the decomposition as a whole.

1. If there's no measurable dissociation between coherence and repertoire status – if constructions diagnosed as low-C_t (repertoire exclusion) systematically exhibit the same interpretive-dispersion profile as constructions diagnosed as low-K, and if tasks designed to separate these signatures fail across a range of phenomena – then the K/C distinction isn't empirically supported.

2. If opportunity-normalized absence doesn't discriminate stable gaps from rarity – if constructions widely treated as "categorical" don't show strong opportunity proxies via large competitor counts (N^*) while rare/uncertain constructions do – then the central methodological claim about opportunity-sensitive negative evidence is undermined.

3. If norm-centre and stakes manipulations don't affect the predicted targets, the conditioning architecture is mis-specified. The $S/A/I$ decomposition yields a specific experimental toolkit: manipulate S via genre/register framing, institutional roleplay, or audience-design cues; manipulate A via speaker ascription cues (biographical metadata, voice/ethnolectal markers, explicitly stated background); manipulate I via explicit norm-centre orientation cues ("speaking as a member of X community", "aiming for formal norms", "in-group banter"); manipulate stakes via consequence framing that should shift $\tau(c)$ without necessarily shifting C_t. If such manipulations don't systematically shift threshold behaviour (as indexed by anchor sets) and don't preferentially affect constructions hypothesized to be repertoire-sensitive, the $S/A/I$ structure isn't just a conceptual repackaging; it's a design toolkit that tells you what counts as a clean manipulation of c rather than a vague "context effect".

4. If the combination rule makes the wrong interaction predictions – if factorial manipulations that independently target coherence and repertoire status show no compounding interaction where the product rule predicts one – then either a different non-compensatory operator is required (e.g. min), or the assumption that the components contribute independently to objective stability is incorrect.

5. If there are robust cases of categorical exclusion with high independent evidence of repertoire membership – if a construction is demonstrably used productively in the relevant c (high production probability and opportunity-normalized corpus rates) and yields stable construals, but is still treated as categorically ungrammatical in repertoire-membership tasks by the same population – then the proposal that grammatical status is constituted by repertoire status plus coherence plus mapping is incomplete.

These conditions are intentionally stated as empirical profiles rather than as verbal counterexamples, since the point is to align theoretical claims with distinct measurement channels.

9 Conclusion

Looking back, grammaticality has functioned as a foundational organizing notion in theoretical linguistics, but it has been burdened with incompatible tasks: marking structural crash, signalling coherence failure, recording community norms, and reporting subjective ungrammaticality. The resulting conceptual overloading has fuelled recurring disputes about whether data is "competence", "performance", or "usage".

Moving forward, grammaticality can be reconceptualized as a state property: conditioned stability of form–value relations within a communicative situation. A minimal decomposition into mapping viability map, interpretive coherence K, and repertoire status C_t yields a compact diagnostic typology and clarifies why acceptability ratings are an imperfect thermometer: not because ratings are useless, but because they're also readings of repair cost, surprisal, and ideology.

The same framework reframes categorical exclusions as potentially emergent stable repertoire exclusion sustained by opportunity-sensitive preemption, and it motivates a concrete research agenda: operationalizing conditioning states, defining the objects of repertoire membership, measuring opportunity sets, and building convergent estimators for repertoire status that don't collapse grammatical status into subjective affect.

Nothing in the formal apparatus is intrinsically linguistic: conditioned stability, repertoire membership, and opportunity-relative measurement apply wherever normed systems produce categorical boundaries. Musical grammar and moral norms are two domains where analogous models may prove productive – both have their own equivalents of asterisks, even when they're not printed.

If grammaticality is to remain a useful concept for theoretical linguistics, it has to become a target of explanation rather than a presupposed label: a hypothesis that can be diagnosed, not an exclamation that ends discussion. The state theory proposed here is intended as a step in that direction: it doesn't replace existing insights about structure, meaning, processing, or norms, but is a minimal architecture that makes their interaction explicit and testable. In this, the asterisk is de-idealized: it stops being a stamp of abstract ill-formedness and becomes a realist diagnostic of stability failure in a situated communicative state.

Acknowledgements

Thanks to Peter Evans, Geoff Pullum, Muhammad Ali Khalidi, Ryan Nefdt, Irene Kosmas, Mostafa Hasrati, and Henri Kauhanen for comments and suggestions.

AI assistance disclosure. The following large language models served as drafting and editing aids throughout the preparation of this paper (January 2026): Claude 3.5 Sonnet and Claude Opus 4.5 (Anthropic, via API and Claude Code CLI); ChatGPT o1 pro and GPT-5.2 pro (OpenAI, via web interface and API); Gemini 3 (Google, via CLI); DeepSeek V3 (DeepSeek, via API). These tools were used for prose drafting, literature search suggestions, LaTeX formatting, and editorial feedback on argumentation. All theoretical claims, arguments, analytical decisions, and interpretive choices are the author's sole responsibility; the models were not used to generate empirical data or to make substantive theoretical determinations.

References

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Krämer & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Royal Netherlands Academy of Arts; Sciences.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733. https://doi.org/10.1353/lan.2006.0186

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Chomsky, N. (1957). *Syntactic structures*. Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley-Interscience. https://doi.org/10.1002/047174882X

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, *41*, 87–100. https://doi.org/10.1146/annurev-anthro-092611-145828

Fanselow, G. (2021). Acceptability, grammar, and processing. In G. Goodall (Ed.), *The cambridge handbook of experimental syntax* (pp. 118–153). Cambridge University Press. https://doi.org/10.1017/9781108569620.006

Featherston, S. (2005). Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua*, *115*(11), 1525–1550. https://doi.org/10.1016/j.lingua.2004.07.003

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). MIT Press.

Gibson, E. A. F. (2026). *Syntax: A cognitive approach*. MIT Press.

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

Goldberg, A. E. (2011). Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics*, *22*(1), 131–153. https://doi.org/10.1515/cogl.2011.006

Grodner, D. J., & Gibson, E. A. F. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*(2), 261–290. https://doi.org/10.1207/s15516709cog0000_7

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge University Press.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

Lakoff, G. (1971). On generative semantics. In D. D. Steinberg & L. A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics, and psychology* (pp. 232–296). Cambridge University Press.

Legendre, G., Grimshaw, J., & Vikner, S. (Eds.). (2001). *Optimality-theoretic syntax*. MIT Press.

Manning, C. D. (2003). Probabilistic syntax. In R. Bod, J. Hay & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 289–341). MIT Press.

McCawley, J. D. (1968). The role of semantics in a grammar. In E. Bach & R. T. Harms (Eds.), *Universals in linguistic theory* (pp. 124–169). Holt, Rinehart; Winston.

O'Connor, C. (2019). Games and kinds. *The British Journal for the Philosophy of Science*, *70*(3), 719–745. https://doi.org/10.1093/bjps/axx027

Pullum, G. K. (2009). More people than you think will understand. Retrieved December 5, 2024, from https://languagelog.ldc.upenn.edu/nll/?p=1997

Pullum, G. K. (2019). Formalism, grammatical rules, and normativity. In J. McElvenny (Ed.), *Form and formalism in linguistics* (pp. 197–223). Language Science Press. https://doi.org/10.5281/zenodo.2654375

Reynolds, B. (2024). *Not every Whose down in Who-ville likes appearing a lot: Pragmatic constraints on independent relative Whose* [Manuscript, LingBuzz/008313]. https://ling.auf.net/lingbuzz/008313

Reynolds, B. (2025). *Words that won't hold still: How linguistic categories work* [Manuscript in preparation]. https://github.com/BrettRey/hpc-book

Reynolds, B. (2026a). *Varieties as conditioning structure: A game-theoretic and Bayesian framework* [Manuscript].

Reynolds, B. (2026b). *Why clause structure is judged like tense and agreement: A coordination account of grammaticality* [LingBuzz 009706]. https://ling.auf.net/lingbuzz/009706

Reynolds, B. (2026c). *Why English doesn't extract left branches (yet)* [LingBuzz 009708]. https://ling.auf.net/lingbuzz/009708

Ritchie, G. D., & Thompson, H. S. (1984). Natural language processing. In T. O'Shea & M. Eisenstadt (Eds.), *Artificial intelligence: Tools, techniques and applications* (pp. 358–388). Harper; Row.

Saussure, F. d. (1916). *Cours de linguistique générale* [Edited by Charles Bally and Albert Sechehaye. English translation: *Course in General Linguistics*, trans. Wade Baskin, New York: Philosophical Library, 1959]. Payot.

Schütze, C. T. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press. https://doi.org/10.17169/langsci.b89.100

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3–4), 379–423, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Silverstein, M. (1976). Shifters, linguistic categories, and cultural description. In K. H. Basso & H. A. Selby (Eds.), *Meaning in anthropology* (pp. 11–55). University of New Mexico Press.

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

Snyder, W. (2022). On the nature of syntactic satiation. *Languages*, *7*(1), 38. https://doi.org/10.3390/languages7010038

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua*, *134*, 219–248. https://doi.org/10.1016/j.lingua.2013.07.002

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237.

Wiese, H. (2023). *Grammatical systems without language borders: Lessons from free-range language*. Language Science Press. https://doi.org/10.5281/zenodo.10276182