Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

One of the Above, None of the Above: Why the Same Model Should Not Predict Both

Lukasz Konowalek

lukasz.konowalek@wum.edu.pl

Medical University of Warsaw

Research Article

Keywords: novelty, category learning, coreference, Pitman–Yor process, Stirling numbers, context

Posted Date: August 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10422267/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

One of the Above, None of the Above: Why the Same Model Should Not Predict Both

Lukasz Konowalek[a,1,*]

[a]*Department of Child Psychiatry, Medical University of Warsaw, Warsaw, Poland*

Abstract

Novelty is often modeled as the probability that the next observation opens a new category. However, experimental setups often leave implicit the fact that novelty judgements are performed within a context that makes such judgements meaningful.

This paper shows that context is necessary for novelty prediction: removing it degrades both combinatorial and Bayesian predictors, and a context-only model with no rich-get-richer term outperforms HG–CRP, the best-performing model in Navarro and Kemp's own six-model comparison, when the task explicitly gates category novelty. We then test whether, within a well-specified context, a combinatorial null model is sufficient for novelty prediction without rich-get-richer effects – while showing that the exact functional form of this partition score is largely interchangeable with simpler density measures in most streams, though not in at least one low-count stream, where the source of its advantage is only partly diagnosed. The results show a nuanced picture: absorption and opening are distinct regimes, but the

*Corresponding author.

Email address: lukasz.konowalek@wum.edu.pl (Lukasz Konowalek)

[1]ORCID: https://orcid.org/0000-0002-4995-680X

direction of the split depends on the local stream's structure. Open, heavy-tailed identity inventories favor the combinatorial score on absorption and Pitman–Yor on opening, whereas small or closed inventories with strong local persistence reverse this pattern. Together with a crossed pattern in a controlled category-learning experiment, this argues against treating category opening as the residual of a single absorption process. Future models of novelty should explicitly separate absorption and opening, and should treat the combinatorial partitioning score as a fixed-context null model whose failures are themselves informative about local stream structure.
Keywords: novelty, category learning, coreference, Pitman–Yor process, Stirling numbers, context

---

1. Introduction

Novelty detection is often cast as a binary partition-extension problem (Pimentel et al., 2014). The next observation either belongs to one of the categories already available in the current context, or it requires a new category. In the category-learning setting of Navarro and Kemp (2017), these alternatives appear as "one of the above" versus "none of the above" responses. Species-sampling models formalize the same contrast: after N observations have occupied K categories, the next observation either joins an existing block or opens a new one (Pitman, 2006; Teh, 2006).

However, this arithmetical binary formulation obscures the epistemic asymmetry between absorption and opening. "One of the above" is an absorption judgement: it asks whether the current item can be resolved within an already generated partition. "None of the above" is an opening judge-

ment: it asks whether the active context licenses a new response identity. Treating opening as merely the complement of absorption risks collapsing these two dependencies into a single probability.

A second problem is that Bayesian partition models begin after the task has already been framed as categorization. They can say how a learner should update beliefs over partitions, but they do not specify how the learner knows that partition extension is the relevant operation in the first place. This is a basic induction problem at the level of the hypothesis space: the same observations can support different inferences under different sampling assumptions (Navarro et al., 2012). From a short history containing only one observed category, the data alone do not decide whether a discrepant item should be absorbed as within-category variation or used to open a new category. Simplicity-based accounts offer one non-Bayesian way of making hypothesis-space constraints explicit (Pothos & Chater, 2002). Opening therefore requires more than counts over the current partition: it requires a context that licenses new response identities.

The analyses below turn this general issue into two empirical questions. First, I ask whether context is necessary for novelty prediction. Here, context is operationalized in a broad dependency-logic-inspired formalization as a local stream unit, an equivalence relation over response identities, and a task-protocol inventory size (Väänänen, 2007). Second, I ask what remains once context is fixed. In open contexts, where there is no fixed category space, the diagnostic null model is combinatorial: novelty is constrained only by the number of compatible partition extensions. In closed contexts, where the category space is fixed in advance, rich-get-richer information,

operationalized here by the Pitman–Yor process, begins to matter. Critically, the winning model reverses between these two judgements within the same stream: the score that best predicts absorption is not the score that best predicts opening, a crossed pattern that is harder to attribute to either model being generically stronger than the other.

I also show that the exact combinatorial score is not merely interchangeable with simpler density surrogates: its advantage in low-count streams is real, though not fully explained by finite-N correction alone, as later sections make clear. Its deviations from density approximations should therefore be treated as empirical clues about local stream structure, consistent with accounts of structured and correlated novelty (Tria et al., 2014, 2018).

2. Methods

The methods section has three jobs. It first defines the dependency atom that turns each corpus into comparable prediction rows; it then defines the block-size-blind combinatorial null and the rich-get-richer comparator; finally, it specifies how held-out losses and uncertainty intervals are computed.

2.1. Episode Streams

Each dataset was converted into an ordered episode stream, indexed by episode t. The construction uses a finite dependency schema over episode streams: each quantity is specified together with the information on which it is allowed to depend. The schema is dependency-logic inspired, but it is not a full team semantics or proof system. Its role is to make feature construction explicit and leakage-safe. The empirical quantities N_t, K_t, and z_t are prefix-dependent rather than item-intrinsic, because they are defined relative to the

preceding history and to a dataset-specific equivalence relation over response identities. This also separates observed partition size from protocol structure: K_t is a prefix statistic, whereas κ_{task} is fixed by the response protocol before the stream begins. Splits are fixed upstream by deterministic hashing of the local stream unit (document, video, or conversation) and, where applicable, episode identity, so that all compared predictors use the same train and test rows within a dataset. Section 2.2 defines these dependencies formally.

For the Navarro–Kemp category-learning experiments, each trial is already a category-novelty judgment rather than a naturalistic recurrence event; Section 2.2 describes how this case differs structurally from the naturalistic streams. Those data are used for the context thesis and for a separate condition-level analogue of the absorption/opening contrast, not for the row-level corpus decomposition.

Naturalistic model comparisons use Bernoulli cross-entropy on held-out test rows. Paired uncertainty intervals are document-cluster bootstraps for GUM documents, Kinetics videos, and SWDA conversations. The Navarro–Kemp reanalysis instead uses condition-level Bernoulli losses and condition-level paired bootstraps, where reported.

The SWDA directional tests were specified before fitting the x_{comb} and Pitman–Yor predictors to SWDA outcomes. The specification used only structural diagnostics of the streams: final inventory size, block-size concentration, and immediate-repetition patterns. Dialogue acts were predicted to be Kinetics-like, while lexical recurrence was predicted to be GUM-like. Table 1 summarizes the confirmatory or exploratory status of every contrast reported below.

[Table 1 about here]

Table 1: Confirmatory and exploratory status of the main contrasts. "Specified before outcome fitting" means that the directional prediction was written from structural diagnostics before fitting x_{comb} or Pitman–Yor to the corresponding outcomes.

Analysis	Status	Basis in this paper
Kinetics-GEBD operator contrast	Exploratory corpus contrast	Included in the original naturalistic comparison; no preregistered direction is claimed.
GUM coreference operator contrast	Exploratory corpus contrast	Included in the original naturalistic comparison; no preregistered direction is claimed.
GUM lexical recurrence operator contrast	Exploratory corpus contrast	Included in the original naturalistic comparison; no preregistered direction is claimed.
SWDA dialogue acts	Specified before outcome fitting	Predicted to be Kinetics-like from finite inventory size and strong immediate repetition.
SWDA lexical recurrence	Specified before outcome fitting	Predicted to be GUM-like from open vocabulary and weaker immediate repetition.
Navarro–Kemp coherence check	Post-hoc coherence check	Noted after inspecting the condition-level result; models and units are only loosely analogous to the corpus split.

2.2. Datasets, Equivalence Relations, and the Dependency Atom

The naturalistic analyses use five episode-stream operationalizations. Let D denote one such operationalization, and let D_{test} denote the held-out episode indices assigned to its test split. Each operationalization defines a finite team of prediction rows, $X_D = \{a_t : t \in D_{\text{test}}\}$, whose elements are dependency atoms:

$$a_t = (C_u, H_{<t}, r_t, z_t), \qquad C_u = (u, \mathcal{R}_u, \kappa_{\text{task}}).$$

Here C_u is the task or annotation context. It contains the local stream unit u, the equivalence relation $\mathcal{R}_u$ that defines response identity inside that unit, and the task-protocol inventory size κ_{task}. The history component $H_{<t}$ is the prefix of earlier episodes in the same local unit, and r_t is the current response identity. This notation keeps the two sources of information separate: κ_{task} is protocol-level context, available before any episode in the stream is observed, while $H_{<t}$ is the empirical prefix available before the current episode is resolved.

The schema contains three dependencies. First, the context and prefix determine the row-level partition statistics, where u_t denotes the local stream unit containing episode t (the same u as in C_u):

$$(C_u, H_{<t}) \Rightarrow (N_t, K_t),$$

with

$$N_t = |H_{<t}|, \qquad K_t = \Big|\{[r_s]_{\mathcal{R}_u} : s < t,\ u_s = u_t\}\Big|.$$

Second, novelty status is determined only after the current response identity is compared with the prefix:

$$(C_u, H_{<t}, r_t) \Rightarrow z_t,$$

where

$$z_t = \mathbf{1}\{\nexists s < t \text{ with } u_s = u_t \text{ and } r_t \,\mathcal{R}_u\, r_s\}.$$

Thus $z_t = 1$ is opening and $z_t = 0$ is absorption. For absorption rows, the current response identity already exists in the prefix; for opening rows, it is first observed at t.

Third, the absorption operator is a protocol-level dependency:

$$C_u \Rightarrow T_{\text{abs}}, \qquad T_{\text{abs}}(a_t) = T_{\text{abs}}(\kappa_{\text{task}}) = \begin{cases} \text{PY}, & \kappa_{\text{task}} < \infty, \\ x_{\text{comb}}, & \kappa_{\text{task}} = \infty. \end{cases}$$

This is not a fitted probabilistic model for z_t. It is a protocol-level selection rule for absorptive rows: finite response inventories predict that absorptive reuse is better captured by rich-get-richer occupancy, whereas open identity spaces predict that absorptive reuse is better captured by the block-size-blind combinatorial score. Opening is reported separately because the evidence does not support this same binary rule there.

For prediction, an admissible naturalistic score must be a function of context and prefix-visible information only:

$$\widehat{p}_t = f(C_u, H_{<t}).$$

The current response identity r_t and novelty label z_t are used to define and evaluate the row, not as inputs to the predictor. Thus κ_{task} is allowed as a context component of the atom, but it is not computed from the current response and does not leak the value of z_t. Table 2 instantiates the dependency context C_u for each of the five naturalistic datasets.

[Table 2 about here]

Table 2: Dataset-specific instantiations of the dependency context $C_u = (u, \mathcal{R}_u, \kappa_{\text{task}})$. The task-protocol inventory size κ_{task} is a property of the response protocol, not the observed number of occupied classes in a stream prefix.

Dataset	Unit u	Equivalence relation $\mathcal{R}_u$	κ_{task}	Source / note
GUM coreference	document	Same annotated coreference chain	∞	Zeldes (2017); no protocol upper bound on document-level chains.
GUM lexical recurrence	document	Same lexical type	∞	Open vocabulary; lexical types are not drawn from a fixed response inventory.
SWDA lexical recurrence	conversation	Same lexical type	∞	Open vocabulary within conversations.
SWDA dialogue acts	conversation	Same SWBD-DAMSL dialogue-act tag	42	Closed dialogue-act inventory (Jurafsky et al., 1997; Stolcke et al., 2000).
Kinetics-GEBD	video	Same perceptual-boundary change label	15	Transition-type labels such as cut, pan, actor change, or action change; not action-phase categories (Kay et al., 2017; Shou et al., 2021).

Table 2 fixes the context component of the atom. All predictors below are then restricted to information available from that context and the prefix $H_{<t}$; the current response identity is used only to score whether the row was absorption or opening.

The Navarro–Kemp reanalysis has a different structure. The unit of anal-

ysis is an experimental novelty judgement rather than a naturalistic recurrence event. The equivalence relation over previously observed items is supplied by the experiment's category structure, while an additional gate variable records whether the task context permits an explicit "none of the above" response (Navarro & Kemp, 2017). This gate is not a partition statistic; it is task-level information about whether opening is licensed in the current judgement context. The structural model reported below uses this gate together with undifferentiated-category indicators, a transition indicator,[2] and the combinatorial score x_{comb} (Eq. (1)) computed from the condition-level partition counts rather than episode-prefix counts and masked to be active only in differentiated absorption states ($K \geq 2$ and not a $\tau = 10$ transition) —; it contains no rich-get-richer block-size term. For the condition-level response decomposition, each condition contributes its observed human none-of-the-above proportion to the none-of-the-above loss component and the remaining mass to the one-of-the-above component. This recombines exactly to the condition-level Bernoulli loss, but it is not a raw trial-level row split.

The finite cases therefore combine two ontologically different kinds of closed inventory. SWDA dialogue acts use a state or act vocabulary: the response identity classifies the current utterance. Kinetics-GEBD uses a transition vocabulary: the response identity classifies the perceptual change

[2]In the implemented reanalysis, the undifferentiated-category indicators mark later $K = 1$ conditions after the initial $N = 1, K = 1$ task-opening case, with an additional $1/N$ term; the transition indicator is the precomputed $\tau = 10$ transition flag used in the Navarro–Kemp condition table.

that produced the current boundary. The operator test deliberately treats both as finite κ_{task} cases. The claim is not that dialogue acts and GEBD labels are the same kind of object, but that protocol-level finiteness predicts the absorptive winner across both kinds.

2.3. Predictors

To quantify rich-get-richer effects, we derived a block-size blind combinatorial score x_{comb}. The combinatorial predictor is computed exactly from log Stirling numbers whenever $K_t \geq 2$ and $K_t \leq N_t$:

$$x_{\text{comb}}(N_t, K_t) = \log S(N_t, K_t - 1) - \log S(N_t, K_t) - \log K_t. \tag{1}$$

This quantity is the finite-count partition-extension score implied by the uniform partition calculation: it asks how the number of admissible partitions changes when the next episode is treated as opening rather than as reuse of an existing identity. It is block-size blind in the sense that it uses only N_t and K_t, not the occupancy distribution of the individual existing response classes. Because $S(N, K)$ weights every set partition of N items into K blocks equally, x_{comb} is a deterministic contrast on a uniformly weighted partition count rather than an exchangeable partition probability function. The Chinese restaurant process and the Pitman–Yor process are size-biased and satisfy the consistency (projectivity) conditions of an exchangeable partition probability function (EPPF); x_{comb} has neither property and is not proposed as a partition prior. Below, it is used purely as a calibrated predictor of z_t, evaluated against Pitman–Yor on held-out data, rather than as a restriction or reparametrization of the Pitman–Yor family. In the main naturalistic comparisons it is evaluated as a one-parameter logistic predictor with no

intercept unless otherwise stated:

$$p(z_t = 1) = \text{logit}^{-1}(\beta x_{\text{comb}}(N_t, K_t)).$$

The density baselines are included because the Stirling calculation itself predicts that much of this score should be expressible as a simple decreasing function of partition density. Appendix A gives the fixed-density limit

$$x_{\text{comb}}(N, K) = g(\rho) + O(1/N), \qquad \rho = N/K.$$

Over the middle range of densities occupied by the corpus data, this limiting function is nearly indistinguishable from a calibrated logarithmic density curve. If the finite-count correction is small, simple logarithmic density should therefore approximate the combinatorial partition-extension score. Matched-complexity tests also use

$$h(\rho, N) = \log \rho + c/N.$$

The c/N term asks whether the exact Stirling score's finite-N curvature matters beyond this middle-range logarithmic approximation. Table 3 reports the fitted degrees of freedom for each of these functional-form predictors.

[Table 3 about here]

Table 3: Fitted degrees of freedom in the functional-form comparisons. "Score-shape" parameters change the predictor before logistic calibration; "calibration" parameters map the score to Bernoulli probabilities on training rows.

Predictor	Score used	Score-shape parameters	Calibration parameters	Total fitted
Exact x_{comb}	Exact log-Stirling score in Eq. (1)	0	1 no-intercept slope β	1
Calibrated $\log \rho$	$\log(N_t/K_t)$	0	1 no-intercept slope β	1
Finite-size density h	$\log \rho + c/N_t$	1 coefficient c	1 no-intercept slope β	2
Pitman–Yor	$(\theta + dK_t)/(\theta + N_t)$	2 parameters (θ, d)	0	2

The Pitman–Yor opening probability is

$$p_{\text{PY}}(z_t = 1) = \frac{\theta + dK_t}{\theta + N_t},$$

with $\theta > 0$ and $0 \leq d < 1$. Pitman–Yor parameters are fitted jointly on all valid training episodes. The absorption/opening decompositions below do not refit Pitman–Yor separately within regimes; they evaluate the same jointly fitted parameters on $z_t = 0$ and $z_t = 1$ held-out rows.

We use the CRP as a comparator rather than as the null model, because it is not block-size blind. In the Chinese restaurant process, the probability of joining an existing block is proportional to that block's current occupancy, so the model already contains a rich-get-richer bias (Pitman, 2006; Teh, 2006). The null model in the present paper is instead x_{comb}, which removes block-size information and asks what can be predicted from the combinatorics of extending a fixed-context partition alone.

2.4. Use of AI Tools

During preparation and revision of this manuscript, the author used Claude (Anthropic) for text editing, for drafting and refactoring analysis code, and for independently checking algebraic derivations (e.g., the saddle-point asymptotics in Appendix A). All AI-assisted output — text, code, and derivations — was reviewed and verified by the author, who takes full responsibility for the content, code, and analyses reported in this manuscript.

3. Results

The results proceed in the same order as the methods. First, local context is tested by replacing local histories with global ones. Second, the fixed-context combinatorial null is compared with simpler density surrogates. Third, the absorption/opening split tests whether the task-protocol component of the atom selects which predictor wins.

3.1. Thesis 1: Context Is Necessary

Local episode boundaries matter. When local histories are replaced by global corpus histories while the prediction target remains local novelty, both x_{comb} and Pitman–Yor degrade (Table 4).

[Table 4 about here]

Table 4: Local-to-global degradation in held-out cross-entropy. Positive deltas mean that respecting local boundaries improves prediction.

Dataset	Δx_{comb}	ΔPY	Local x_{comb}	Global x_{comb}
Kinetics	0.1656	0.3388	0.4825	0.6481
GUM lexical	0.0382	0.2812	0.6457	0.6839
GUM coreference	0.0069	0.0226	0.6859	0.6927

A density-only control gives the same lesson in sharper form. In Kinetics, replacing local density by global density changes CE from 0.455 to 1.939. The analogous degradation is also present in GUM lexical recurrence (0.638 to 1.883) and GUM coreference (0.667 to 0.693). Thus context is not a decoration around a partition model; it determines which partition is being extended.

The Navarro–Kemp experiments give a related conclusion in a smaller controlled setting. In Experiment 4, a task-opening structural model — combining the task's "none of the above" gate with partition-density and transition features, but no rich-get-richer term — beats the HG–CRP benchmark: CE 0.587 versus 0.710, $\Delta = -0.123$, 95% CI $[-0.192, -0.060]$ (condition-level paired bootstrap). Across the full set of conditions, the same model narrows but does not close this gap (CE 0.608 versus 0.628), consistent with the task gate carrying most of the signal outside Experiment 4's specific contrast.

Together, these tests establish the first dependency claim: novelty prediction is not defined by a partition alone, but by the local context in which that partition is being extended.

3.2. The Fixed-Context Combinatorial Null

Once the local context is fixed, the testable null model is not a free global novelty prior. It is the partition-extension score $x_{\text{comb}}(N, K)$: a finite-count statistic derived from how the number of admissible partitions changes when an episode is treated as opening rather than as absorption. The remaining results ask whether this null model is sufficient, where it fails, and whether those failures require rich-get-richer information.

3.2.1. A Note on Functional Form

The exact functional form of x_{comb} is not itself uniquely predictive in the empirical middle range. Appendix A gives a limiting curve $g(\rho)$, and over the observed GUM density range that curve is closely approximated by a calibrated logarithmic density. On GUM opening episodes, giving x_{comb} and $\log \rho$ the same calibration freedom therefore makes them nearly interchangeable: calibrated $\log \rho$ closes 95.0% of the original x_{comb}–Pitman–Yor opening gap in GUM lexical recurrence and 61.5% in GUM coreference, against 93.9% and 60.8% for calibrated x_{comb} itself. In the all-valid held-out comparison, the generic finite-size density model $h(\rho, N)$ ties or beats exact x_{comb} in both GUM streams. Because h has an additional fitted finite-size shape parameter c (Table 3), this should be read as comparability with a more flexible density surrogate, not as equivalence between the two scores.

The exact Stirling form is most interpretable when the finite-count correction is non-negligible. Appendix A predicts that, at fixed $\rho = N/K$, $x_{\text{comb}}(N, K)$ should converge to $g(\rho)$ with an $O(1/N)$ residual, and the GUM streams behave this way: the observed exponent on N is -0.958 (GUM lexical) and -1.087 (GUM coreference), consistent with the predicted rate.

Kinetics does not show this asymptotic pattern (raw exponent -0.041), and exact x_{comb} does beat the fitted finite-size density surrogate in the all-valid Kinetics comparison (0.4825 vs. 0.5325 CE). However, a post-hoc stratification by N_t does not show that this advantage is confined to larger departures from the asymptotic regime: using the globally fitted h parameters, the $h - x_{\mathrm{comb}}$ gap is 0.0158 CE for $N_t \leq 4$ and 0.0940 CE for $N_t > 4$. Thus the Kinetics result should be read as evidence that the exact finite Stirling score captures something missed by this particular finite-size surrogate, not as a clean monotone demonstration that the advantage is caused by low N_t alone. In short, $\log \rho$ works where the limiting curve is already effectively logarithmic, while exact x_{comb} remains the parameter-free finite-count score; its empirical advantage in Kinetics is real but not isolated by this simple N_t split.

3.3. Thesis 2: Task-Protocol Inventory Classifies the Absorptive Winner

The all-valid aggregate hides two prediction regimes. Absorption rows ($z_t = 0$) ask whether the current episode reuses an existing response identity; opening rows ($z_t = 1$) ask whether it creates a new one. The central operator test asks whether the absorptive winner is selected by $T_{\mathrm{abs}}(a_t)$, which depends on the atom only through its protocol component κ_{task}. Opening is reported alongside absorption, but as a diagnostic comparison rather than as the same binary operator.

Table 5 gives the operator test. For absorption, T_{abs} is consistent with all five observed domain contrasts: finite protocol inventories favor Pitman–Yor, and open inventories favor x_{comb}. This should be read as a compact five-domain test rather than as a large independent sample: the finite side

contains two independent cases (Kinetics-GEBD and SWDA dialogue acts), and the three open-inventory cases are conceptually closer to one another because all involve open lexical or identity recurrence. Opening is reported in the same table, but it is not treated as governed by the same binary rule. Instead it provides a weaker ordering diagnostic: Kinetics-GEBD, the smallest finite-inventory case, favors x_{comb} on opening; the three open-inventory domains favor Pitman–Yor; and SWDA dialogue acts, with a finite but larger inventory of 42 tags, is a boundary case whose interval includes zero.

[Table 5 about here]

Table 5: Task-protocol operator test. The contrast is $\Delta = \text{CE}_{x_{\text{comb}}} - \text{CE}_{\text{PY}}$, so negative values favor x_{comb} and positive values favor Pitman–Yor. Bracketed intervals are 95% cluster bootstraps for Δ. The binary operator T_{abs} is evaluated only for absorption: finite κ_{task} predicts Pitman–Yor and open κ_{task} predicts x_{comb}. Opening is shown as a regime diagnostic rather than as the same binary test.

Dataset	κ_{task}	Source	Abs. Δ	Abs. CI	Open. Δ	Open. CI	Abs. match	Opening status
Kinetics-GEBD	15	Shou et al. (2021)	0.0778	[0.0765, 0.0792]	-0.1655	[-0.1745, -0.1561]	yes	x_{comb}
SWDA dialogue acts	42	Jurafsky et al. (1997) and Stolcke et al. (2000)	0.0221	[0.0175, 0.0266]	0.0326	[-0.0253, 0.0933]	yes	unresolved
GUM coreference	∞	Zeldes (2017)	-0.2190	[-0.2390, -0.2006]	0.2222	[0.2096, 0.2341]	yes	Pitman–Yor
SWDA lexical recurrence	∞	open vocabulary	-0.0165	[-0.0181, -0.0150]	0.0541	[0.0509, 0.0573]	yes	Pitman–Yor
GUM lexical recurrence	∞	open vocabulary	-0.0486	[-0.0507, -0.0466]	0.0945	[0.0918, 0.0973]	yes	Pitman–Yor

Figure 1 visualizes the same result as a protocol-grouped dot plot. Read it as a sign plot: points above zero favor Pitman–Yor, and points below zero favor x_{comb}. The finite-inventory cases are shown separately, and the three open-inventory streams are jittered within a single open-inventory group. The figure makes the opening pattern visible without treating it as the same binary operator: SWDA dialogue-act opening lies closest to zero, consistent with a graded or threshold extension rather than a binary rule for opening.

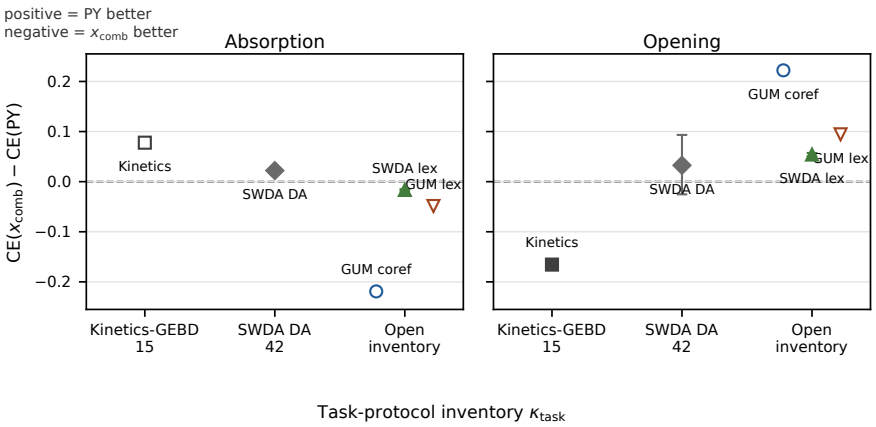


Figure 1: Task-protocol inventory size and regime-specific model advantage. The x-axis groups domains by protocol-level response inventory κ_{task}: two finite inventories and one open-inventory group. The y-axis is $\text{CE}_{x_{\text{comb}}} - \text{CE}_{\text{PY}}$, so positive values favor Pitman–Yor and negative values favor x_{comb}. Absorption is consistent with the finite/open operator in all five domains, with the caveat that the finite side contains only two independent cases. Opening is shown as a weaker diagnostic ordering, with SWDA dialogue acts as the boundary case.

3.3.1. Row-Level Decomposition

Table 6 reports the row-level decomposition for all naturalistic streams. In GUM coreference, GUM lexical recurrence, and preregistered SWDA lexical recurrence, x_{comb} is better on absorption and Pitman–Yor is better on opening. In Kinetics, the direction reverses: Pitman–Yor is better on absorption, while x_{comb} is better on opening. SWDA dialogue acts match the Kinetics-like prediction on absorption, but opening is unresolved: Pitman–Yor is better pointwise, while the bootstrap interval crosses zero.

[Table 6 about here]

Table 6: Absorption/opening decomposition in naturalistic streams. The contrast is $\text{CE}_{x_{\text{comb}}} - \text{CE}_{\text{PY}}$, so negative values favor x_{comb} and positive values favor Pitman–Yor. Row-count-weighted recombination reproduces the all-valid CE values to numerical precision.

Dataset	Subset	n	$\text{CE}_{x_{\text{comb}}}$	CE_{PY}	Δ
GUM coref	All valid	16389	0.6859	0.6776	0.0083
GUM coref	Absorption	7947	0.5972	0.8162	-0.2190
GUM coref	Opening	8442	0.7694	0.5472	0.2222
GUM lexical	All valid	47102	0.6457	0.6355	0.0102
GUM lexical	Absorption	27754	0.4487	0.4973	-0.0486
GUM lexical	Opening	19348	0.9284	0.8339	0.0945
Kinetics	All valid	62144	0.4825	0.4468	0.0357
Kinetics	Absorption	51379	0.2660	0.1882	0.0778
Kinetics	Opening	10765	1.5155	1.6810	-0.1655
SWDA lexical	All valid	238288	0.5950	0.5884	0.0066
SWDA lexical	Absorption	160338	0.3525	0.3690	-0.0165
SWDA lexical	Opening	77950	1.0938	1.0397	0.0541
SWDA DA	All valid	42869	0.2707	0.2477	0.0230
SWDA DA	Absorption	39262	0.1028	0.0807	0.0221
SWDA DA	Opening	3607	2.0983	2.0657	0.0326

The preregistered SWDA test is important because the expected direc-

tions were set before fitting either predictor. SWDA lexical recurrence was predicted to be GUM-like: a large lexical inventory and weak immediate repetition should favor x_{comb} on absorption and Pitman–Yor on opening. Both predictions hold, with bootstrap intervals excluding zero for absorption ($[-0.0181, -0.0150]$) and opening ($[0.0509, 0.0573]$). SWDA dialogue acts were predicted to be Kinetics-like because the inventory is smaller and more autocorrelated. The absorptive prediction holds ($\Delta = 0.0221$, CI from 0.0175 to 0.0266), but the opening side does not resolve in the predicted direction ($\Delta = 0.0326$, CI from -0.0253 to 0.0933).

Figure 2 shows the original GUM coreference interaction directly. The CRP curve is included only as the standard species-sampling reference point between the block-size blind combinatorial null and the discounted Pitman–Yor fit; it is not treated as a separate null model. In that stream, the combinatorial score is best on absorption, where the current episode reuses an existing equivalence class, and the rich-get-richer Pitman–Yor fit is best on opening, where the current episode creates a new class. The expanded table shows that this crossed interaction is one member of a broader family rather than a universal sign pattern.

Because both x_{comb} and Pitman–Yor are functions of (N_t, K_t), this crossed regime interaction could in principle reflect systematic differences in the (N_t, K_t) distribution between absorption and opening rows rather than a regime-specific effect. A post-hoc diagnostic confirms that absorption rows are systematically later and denser than opening rows in every stream. In GUM coreference, however, the crossed pattern survives this control: x_{comb} wins absorption and Pitman–Yor wins opening within every N_t bin, and a

row-level regression controlling for $\log(1 + N_t)$ retains 101.4% of the uncontrolled regime coefficient (104.9% after additionally controlling for $\log(1+K_t)$ and $\log(N_t/K_t)$). The same bin-wise stability holds in GUM lexical recurrence. The pattern is present but less uniformly stable in Kinetics (opening consistent in 3 of 4 N_t bins) and in the two SWDA streams, where bin-wise signs are less stable. For SWDA dialogue acts this qualifies a contrast whose opening side was already unresolved, and for SWDA lexical recurrence it qualifies, without reversing, the preregistered whole-sample direction. The SWDA lexical reconstruction used for this diagnostic was rebuilt from `swda_raw_dialog_acts.parquet` and contains 16 fewer rows than the original Table 6 sample (n = 238,272 vs. 238,288); cross-entropy values match at the reported precision.

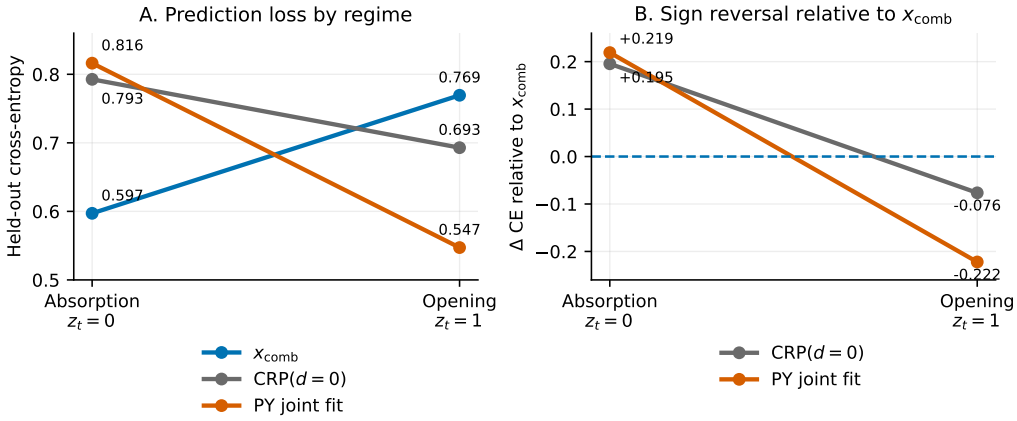


Figure 2: Interaction between combinatorial density and rich-get-richer information in GUM coreference. Panel A shows held-out cross-entropy by regime. Panel B subtracts x_{comb} from each model's loss, so positive values mean worse than the combinatorial score and negative values mean better. CRP is shown as the standard species-sampling comparator, not as the paper's null model. In this open, heavy-tailed stream, x_{comb} is favored on absorption while the joint Pitman–Yor fit is favored on opening.

The all-valid value is not a separate effect: recombining the absorption and opening losses by their row counts reproduces the aggregate to numerical precision in each corpus. The aggregate comparison can therefore conceal large and sometimes oppositely signed regime effects.

3.3.2. Experimental Category Judgements

The Navarro–Kemp data provide a related but methodologically different check. They are available at condition level rather than as individual trial-level choices, so they cannot supply the same row-level held-out decomposition as Table 6. A soft decomposition of the aggregated Bernoulli loss nevertheless separates the "one of the above" and "none of the above" components. In Experiment 4, HG–CRP is better on the one-of-the-above

component (CE 0.165 versus 0.328 for the task-opening structural model), while the task-opening structural model is better on the none-of-the-above component (CE 1.209 versus 2.019 for HG–CRP). This is not pooled with the corpus table because the units and models differ: it is a condition-level gate/HG–CRP comparison, not a row-level x_{comb}/Pitman–Yor split.

Category-learning experiments of this kind typically involve a small, fixed set of experimenter-defined categories per condition, structurally closer to Kinetics and SWDA dialogue acts than to GUM's open lexical or coreference space. In that setting, HG–CRP's advantage on one-of-the-above and the gate's advantage on none-of-the-above match the small-inventory pattern seen in Kinetics and dialogue-act absorption. This alignment was noted after inspecting the result rather than predicted in advance, and it compares models (gate, HG–CRP) that are only loosely analogous to x_{comb} and Pitman–Yor, so it is reported as a coherence check rather than as independent confirmation of the task-protocol operator.

4. Discussion

Traditional accounts model novelty as probability mass that is not assigned to existing categories. The present results show that this is not always the right default assumption. In open identity spaces, the combinatorial null x_{comb} is sufficient to predict absorption, while rich-get-richer Pitman–Yor information is needed to predict opening. In small closed inventories, the pattern reverses: Pitman–Yor is better on absorption, while x_{comb} is better on opening. The same Pitman–Yor term can therefore be useful in finite dialogue-act or boundary-label systems but is not the appropriate default

null model for open lexical or coreference identity spaces. This crossed pattern is consistent with the view that absorption and opening are governed by distinct dependency structures, rather than being two faces of a single novelty mechanism: the one-of-the-above/none-of-the-above contrast should not be treated as a single probability with one generative process behind it.

The discussion draws out three implications of this result: for Bayesian accounts of category learning, for memory-system analogies between integration and differentiation, and for the role of combinatorial scores as null models rather than as complete psychological mechanisms.

4.1. One of the Above vs. None of the Above

The results bear directly on Bayesian accounts of category learning, in which opening a new category is typically the residual of a single partition-inference process: a cluster opens when this gives the best posterior account of the data (Anderson, 1991; Sanborn et al., 2010), and Navarro and Kemp's "one of the above" versus "none of the above" task treats the two responses as two readings of one underlying decision (Navarro & Kemp, 2017). The crossed pattern in Table 6 is not consistent with this residual-event view: if opening were simply the complement of absorption within one process, the same predictor should not systematically win one side and lose the other. The task-protocol operator (Section 3.3) makes this concrete: whether an open-inventory or closed-inventory account wins depends on which side of the contrast is being predicted, not on which model is generically better suited to the domain.

This reading is compatible with multiple-system accounts of category learning, which hold that different categorization demands recruit different

representational and control processes rather than a single mechanism (Ashby & Maddox, 2005, 2011; Ashby et al., 1998) (Section 4.2 extends the point in relation to memory-systems accounts of integration and differentiation). It does not follow, however, that "one of the above" and "none of the above" are always separable in practice: the SWDA dialogue-act opening result was unresolved (Table 6), and the present naturalistic streams are annotation and discourse tasks rather than the controlled category judgements Navarro and Kemp study. What the results support is narrower than a general claim about categorization: after context and response protocol are fixed, the information that best predicts reuse of an existing identity is not reliably the information that best predicts the decision to open a new one.

It is worth being explicit about what the present results say in relation to Navarro and Kemp's own findings, since Section 3.1 already reports a direct comparison against their strongest model. In Experiment 4, a task-opening structural model — combining the "none of the above" gate with partition-density and transition features, but no rich-get-richer term at all — outperforms HG–CRP, the best-performing model in Navarro and Kemp's own comparison across six variants: CE 0.587 versus 0.710, a difference whose 95% interval excludes zero. This is not a marginal or loosely analogous result: it shows that knowing whether the task licenses a novel-category response at all is more informative, in this experiment, than the occupancy-and-uncertainty machinery that Navarro and Kemp built to explain their data. Read alongside Experiment 4's regime decomposition — HG–CRP wins one-of-the-above (CE 0.165 versus 0.328) while the gate model wins none-of-the-above (CE 1.209 versus 2.019) — the same crossed pattern re-

ported for x_{comb} and Pitman–Yor in the naturalistic corpora reappears in a controlled category-learning experiment, using a completely different family of models. Taken together, a controlled laboratory task and five naturalistic corpora agree that no single model wins on both sides of the contrast: the residual-event view of novelty, in which opening is simply whatever probability mass absorption leaves behind, looks like the wrong default rather than a convenient simplification with occasional exceptions.

4.2. A Memory-Systems Analogy

The absorption/opening distinction also resonates with a classic problem in theories of hippocampal memory: how a system can both separate similar inputs into distinct traces and retrieve or complete an existing trace from partial input. Marr's theory of archicortex treated hippocampal memory as a rapid associative system for forming simple traces (Marr, 1971). Later computational models made the separation/completion trade-off explicit: hippocampal conjunctive coding can support rapid storage and recall while avoiding interference between overlapping episodes (O'Reilly & McClelland, 1994). Contemporary reviews describe pattern separation as the process by which similar inputs are made more distinct in hippocampal representations, complementing retrieval processes that recover stored structure from partial cues (Yassa & Stark, 2011).

The present analysis is not a neural model of hippocampus. Its contribution is at a different level: it asks when an observed item should be treated as evidence for reuse of an existing identity and when it should be treated as evidence for opening a new one. Still, the memory-systems literature makes clear why this distinction is not merely terminological. Absorption is anal-

ogous to completion or integration into an existing representation; opening is analogous to separation or creation of a new trace. The empirical result that different predictors win on the two sides of the judgement does not test this memory-systems account, but it is structurally compatible with it: both describe integration and differentiation as operations that can be selectively engaged, rather than as two readouts of a single underlying process.

4.3. Open and Closed Response Systems

The same result can be read as a distinction between open and closed response systems. In a closed response system, the task protocol fixes a finite inventory of possible identities before the stream begins. Reuse is then informative because the learner or model is choosing among a bounded set of alternatives: occupancy and persistence can legitimately carry predictive weight. In an open response system, by contrast, the observed partition is only a prefix of an unbounded identity space. Failure to match a previous identity does not exhaust the possible responses, so the natural fixed-context null model is combinatorial rather than rich-get-richer, consistent with the absorption-side operator test in Table 5.

This distinction is related to, but not identical with, the open-set/closed-set contrast in machine learning and the open-world/closed-world distinction in knowledge representation. Open-set recognition makes explicit that classification changes when test items may belong to classes absent from the training set (Scheirer et al., 2013). Closed-world reasoning, conversely, treats the known database as complete for the predicates at issue (Reiter, 1978). The present analysis applies the same structural contrast to novelty judgements: whether repetition or opening is predictable from occupancy depends

on whether the task has fixed the response universe in advance. This is why the same Pitman–Yor term can be useful in finite dialogue-act or boundary-label systems but is not the appropriate default null model for open lexical or coreference identity spaces.

4.4. The Combinatorial Null

The value of the combinatorial score is diagnostic. Because it ignores the occupancy of particular response identities, it gives a fixed-context baseline for what can be predicted from partition size alone. This is the role of a null model in the present argument: not a claim that the null is psychologically complete, but a deliberately restricted model whose failures identify which additional structure matters (Gotelli & Graves, 1996; Wimsatt, 1987). When Pitman–Yor improves on x_{comb}, the improvement can therefore be read as evidence that identity-specific reuse or persistence matters in that domain, rather than merely as a better curve fit. Conversely, when x_{comb} matches or outperforms the rich-get-richer predictor, the extra occupancy information has not earned its keep once the task context and prefix partition are fixed. GUM coreference illustrates both readings within a single stream: x_{comb} outperforms Pitman–Yor on absorption, where occupancy adds nothing beyond partition size, while Pitman–Yor overtakes x_{comb} on opening, where it does. This reading is most secure where it has been checked against the shared (N_t, K_t) dependence of both scores (Table 6); it is correspondingly more tentative in the streams where that check was less conclusive.

5. Limitations

The naturalistic episode streams are operationalizations of discourse, dialogue, and action annotation, not direct measurements of cognitive states. The response identities used here are lexical types, coreference chains, dialogue-act tags, and annotation labels; other task definitions could produce different partitions.

The Navarro–Kemp reanalysis is diagnostic rather than a full replacement model for the original category-learning experiments. It shows that a task gate can explain much of the relevant contrast, but it does not model every aspect of human similarity judgment.

We attempted to test whether the absorption/opening distinction found in GUM also structures the Navarro–Kemp category judgements directly, analogous to Table 6. This was not possible at the row level: the available Navarro–Kemp data are aggregated at the condition level rather than recorded as individual trial-level choices. A soft decomposition of the aggregated Bernoulli loss into its "one of the above" and "none of the above" components shows an asymmetry between the gate and HG–CRP models that is consistent with the small-inventory pattern seen in Kinetics and SWDA dialogue acts, though this comparison is not equivalent to an independent held-out row split and was not preregistered in this form. Whether the absorption/opening distinction generalizes to explicit human categorization judgements therefore remains an open question requiring trial-level data.

The Experiment 4 comparison against HG–CRP (Section 4, and Section 4.1 in the Discussion) has not been subjected to the same (N_t, K_t)-style robustness check applied to the naturalistic streams below. The gate model's

advantage on the none-of-the-above component and HG–CRP's advantage on one-of-the-above are computed from condition-level aggregates rather than trial-level rows, so a bin-wise or regression-based control analogous to the one used for GUM coreference is not straightforwardly available with the data as released. The asymmetry is large and its 95% interval excludes zero at the level of the whole-experiment comparison, but whether it would survive a control for, say, the number of displayed categories or trials per condition has not been tested. We flag this as the least independently verified of the crossed-pattern results emphasized in the Discussion, notwithstanding its size.

The SWDA dialogue-act opening result did not resolve in the predicted direction, and an exploratory (non-preregistered) follow-up found that the per-conversation loss difference is widely and roughly symmetrically distributed (54.4% of conversations favoring x_{comb}, 45.6% favoring Pitman–Yor), consistent with genuine between-conversation heterogeneity rather than an underpowered test of a fixed effect. This heterogeneity correlates with conversation-level block-size statistics, and the association remains after controlling for conversation-level mean N_t (for mean block size, partial Spearman $\rho = 0.54$). However, because x_{comb} and Pitman–Yor are themselves functions of (N_t, K_t), part of this correlation may still reflect the mechanical sensitivity of the two scores to different (N_t, K_t) trajectories rather than an independent generative effect. We treat this as an open question rather than a further test of the task-protocol operator.

The crossed absorption/opening interaction in Table 6 also rests on models that are themselves functions of (N_t, K_t), and absorption rows are system-

atically later and denser than opening rows in all five streams. A bin-wise and regression-based check leaves the GUM coreference and GUM lexical results essentially unchanged, with 101–105% of the uncontrolled regime effect retained in GUM coreference after N_t/K_t controls. The check is less clean in SWDA lexical recurrence, where the preregistered whole-sample direction holds but bin-wise signs are less stable across N_t, and in Kinetics opening, where the expected direction appears in 3 of 4 bins. This does not overturn the whole-sample results in Table 6, but it means the crossed-interaction interpretation is best supported in the two open-inventory GUM streams and should be read with more caution elsewhere.

A related but distinct robustness question concerns the Kinetics matched-complexity comparison in Section 3.2.1, rather than the crossed-regime interaction discussed above. Although exact x_{comb} outperforms the fitted $h(\rho, N) = \log \rho + c/N$ surrogate on all valid Kinetics rows, splitting the same held-out comparison by N_t does not make the advantage attenuate with larger N_t. This weakens the simple explanation that the Kinetics result is driven only by low-count finite-size correction. The safer conclusion is that x_{comb} is a parameter-free finite-count baseline that outperforms this particular density-plus-$1/N$ surrogate in Kinetics; identifying which aspect of the exact Stirling form matters there requires a more targeted analysis.

Two further methodological limitations concern the comparisons reported above rather than any single number in them. First, the Navarro–Kemp task-opening structural model is a discriminative regression over the task gate, partition density, and transition features, benchmarked against Navarro and Kemp's own published HG–CRP predictions; it is not a re-fitted Bayesian

mixture model in which a partition prior and a similarity likelihood are explicitly separated. It can therefore show that task-context information is more predictive than HG–CRP's occupancy-and-uncertainty machinery as a whole, but it cannot by itself attribute HG–CRP's disadvantage to the prior component, the likelihood component, or their interaction. Second, although every predictive contrast above is reported with cluster-bootstrap uncertainty, the shape parameters of the finite-size density surrogate (c in $h(\rho, N)$) and of the jointly fitted Pitman–Yor process (θ, d) do not have a dedicated recovery or identifiability study in this paper. Stability of the Pitman–Yor fit across a small grid of optimizer restarts was checked (Supplement), but formal identifiability, particularly of c under short streams, was not.

All naturalistic and Navarro–Kemp comparisons above use held-out cross-entropy, a proper scoring rule that jointly reflects calibration and discrimination. We do not separately report AUC or binned calibration curves in this paper; two predictors that are indistinguishable on cross-entropy could still differ in rank-based discrimination or in bin-wise calibration, and that possibility is not ruled out by the analyses above.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The author declares no competing interests.

Ethics approval and consent to participate

This study reanalyzes previously published, publicly available corpora (GUM, SWDA, and Kinetics-400 event-boundary annotations) and previously published, aggregated condition-level results from Navarro and Kemp (2017). No new data were collected from human participants by the author, and no ethics approval was required for this study.

Consent to publish

Not applicable; this study did not involve the collection of new data from identifiable human participants.

Data availability

The GUM corpus is available at https://github.com/amir-zeldes/gum. The Kinetics-400 dataset is available at https://github.com/cvdfoundation/kinetics-dataset; the Kinetics-GEBD event-boundary annotations (Shou et al., 2021) are available at https://github.com/StanLei52/GEBD. The Switchboard Dialog Act corpus is available at https://web.stanford.edu/~jurafsky/ws97. The Navarro–Kemp (2017) condition-level data are as published in the cited article. The derived episode-level and condition-level datasets used in the analyses above, together with the generated tables and figures, are archived at Open Science Framework, https://doi.org/10.17605/OSF.IO/MCG82, and are also provided as a supplementary ZIP archive with this submission.

Code availability

All analysis code is archived at Open Science Framework, https://doi.org/10.17605/OSF.IO/MCG82, and is also provided as a supplementary ZIP

archive with this submission. The archive includes a reconstruction of the script that generates the SWDA rows of Table 6, since the original script that produced these results could not be recovered; the reconstruction matches the reported values closely but not exactly, and the discrepancy is documented in the Supplement (Section 6.2).

Author contributions

Lukasz Konowalek: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review and editing.

Acknowledgements

The author thanks Nicholas Chater and Thomas L. Griffiths for helpful comments that improved the clarity of the manuscript, and Vadim Weinstein for pointing out the analogy with dependence logic.

Appendix A. Saddle–Point Asymptotics of the Stirling Numbers

Throughout the appendix we work with $\nu = K/N = 1/\rho$ for convenience.

From the generating function $(e^x - 1)^K/K! = \sum_{N \geq K} S(N,K)\, x^N/N!$ one obtains the contour integral representation

$$S(N,K) = \frac{N!}{K!}\frac{1}{2\pi i}\int_{\mathcal{C}} \exp\Big(\phi(x)\Big)\frac{dx}{x}, \qquad \phi(x) = -N\log x + K\log(e^x - 1).$$

The saddle point $x_0(\nu)$ is the unique positive solution of

$$\nu = \frac{1 - e^{-x_0}}{x_0}, \tag{A.1}$$

which is strictly decreasing in $\nu \in (0,1)$.

Taking logarithms of the one-term saddle-point approximation of Temme, 1993, eq. 2.9 gives

$$\log S(N,K) = N\Phi(\nu) + C(N,K) + O\Big(\frac{1}{N}\Big), \tag{A.2}$$

where the rate function is

$$\Phi(\nu) = -\log x_0(\nu) + \nu \log(e^{x_0(\nu)} - 1), \tag{A.3}$$

and the sublinear correction term is

$$\begin{aligned} C(N,K) = \log \frac{N!}{K!(N-K)!} + \log(N-K)! - \tfrac{1}{2}\log\frac{(N-K)}{2\pi} \\ + \log f(t_0) + O\Big(\frac{1}{N}\Big), \end{aligned} \tag{A.4}$$

where the first two terms combine exactly, yielding the more compact form used in the derivation:

$$C(N,K) = \log \frac{N!}{K!} + \tfrac{1}{2}\log\frac{2\pi}{N-K} + \log f(t_0) + O\Big(\frac{1}{N}\Big). \tag{A.5}$$

Here $t_0 = (N-K)/K$, and

$$f(t_0) = \sqrt{\frac{t_0}{(1+t_0)(x_0 - t_0)}},$$

which is real and positive since $x_0 > t_0$ holds for all $\nu \in (0,1)$, as follows directly from the saddle equation (A.1) and $t_0 = (1-\nu)/\nu$.

The approximation (A.2) is uniformly valid in ν bounded away from 0 and 1; relative errors for moderate N are reported in Table A.7.

Table A.7: Relative error of the Temme approximation (A.2) for $\log S(N, K)$.

N	K	$\log S(N,K)$ (exact)	Relative error
20	5	27.342	1.0×10^{-4}
20	10	29.409	8.6×10^{-5}
50	15	106.991	1.1×10^{-5}
50	25	98.717	1.0×10^{-5}
100	30	264.367	2.3×10^{-6}
100	50	234.022	2.1×10^{-6}
200	60	627.981	4.8×10^{-7}

Let $\psi(x;\nu) = \phi(x)/N = -\log x + \nu \log(e^x - 1)$ denote the normalized phase function, so that the saddle equation is $\partial_x \psi(x_0;\nu) = 0$. By the envelope theorem,

$$\Phi'(\nu) = \left.\frac{\partial \psi}{\partial \nu}\right|_{x=x_0} = \log(e^{x_0(\nu)} - 1). \tag{A.6}$$

Appendix A.1. Asymptotic Logit

We work in the joint asymptotic regime $N, K \to \infty$ with $\nu = K/N$ held fixed. The derivation proceeds in three steps.

Step 1: difference of the leading term. By a first-order Taylor expansion of Φ in ν,

$$N\Phi\Big(\nu - \tfrac{1}{N}\Big) - N\Phi(\nu) = -\Phi'(\nu) + O\Big(\frac{1}{N}\Big) = -\log(e^{x_0(\nu)} - 1) + O\Big(\frac{1}{N}\Big), \tag{A.7}$$

where the $O(1/N)$ remainder is $\Phi''(\nu)/(2N) + O(1/N^2)$.

Step 2: difference of the correction term. The dominant contribution to $C(N,K)$ in (A.4) is $\log(N!/K!)$. When K decreases by one,

$$\log \frac{N!}{(K-1)!} - \log \frac{N!}{K!} = \log K.$$

The remaining terms in $C(N,K)$ change by $O(1/N)$: the term $\frac{1}{2}\log(2\pi/(N-K))$ changes by $-1/(2(N-K)) = O(1/N)$, which requires $N-K = \Theta(N)$, i.e. ν bounded away from 1, and $\log f(t_0)$ changes by $O(1/N)$ since t_0 and x_0 vary continuously with ν. Therefore,

$$C(N,K-1) - C(N,K) = \log K + O\Big(\frac{1}{N}\Big). \tag{A.8}$$

Table A.8 confirms that the residual $[C(N,K-1) - C(N,K)] - \log K$ is $O(1/N)$.

Table A.8: Verification of (A.8): the residual $[C(N,K-1) - C(N,K)] - \log K$ converges to zero at rate $O(1/N)$. Here $C(N,K-1) - C(N,K)$ is computed as $[\log S(N,K-1) - \log S(N,K)] - [N\Phi((K-1)/N) - N\Phi(K/N)]$.

ν	N	K	$C(N,K-1) - C(N,K)$	$\log K$	Residual $\times N$
0.2	100	20	2.99225	2.99573	−0.348
0.2	200	40	3.68707	3.68888	−0.362
0.2	500	100	4.60443	4.60517	−0.369
0.2	1000	200	5.29795	5.29832	−0.372
0.2	2000	400	5.99128	5.99146	−0.373
0.3	100	30	3.39425	3.40120	−0.695
0.3	200	60	4.09083	4.09434	−0.703
0.3	500	150	5.00922	5.01064	−0.707
0.3	1000	300	5.70307	5.70378	−0.709
0.3	2000	600	6.39657	6.39693	−0.710

Step 3: assembling the structural score. Combining Steps 1 and 2,

$$\log S(N, K-1) - \log S(N, K) = -\log(e^{x_0(\nu)} - 1) + \log K + O\Big(\frac{1}{N}\Big).$$

Substituting into the exact score (1),

$$\begin{aligned} x_{\text{comb}}(N, K) &= \Big[\log S(N, K-1) - \log S(N, K)\Big] - \log K \\ &= -\log(e^{x_0(\nu)} - 1) + \log K - \log K + O\Big(\frac{1}{N}\Big). \end{aligned}$$

The two $\log K$ terms cancel exactly, leaving

$$x_{\text{comb}}(N, K) = g(\rho) + O\Big(\frac{1}{N}\Big), \qquad g(\rho) = -\log\Big(e^{x_0(1/\rho)} - 1\Big), \tag{A.9}$$

where $\rho = N/K$. The cancellation is exact: the $+\log K$ in the difference of log-Stirling numbers comes from the $\log(N!/K!)$ prefactor in Temme's formula, and the $-\log K$ comes from the definition of the score in (1).

Intuitively, $g(\rho)$ is the smooth density-only curve that remains after the finite partition count is sent to its fixed-density limit. The exact Stirling score differs from this curve by the finite-N correction, which is why low-count streams can retain curvature that a simple density surrogate does not capture.

Table A.9 confirms that $N \times (x_{\text{comb}} - g(\rho))$ converges to a finite constant, establishing the $O(1/N)$ rate empirically.

Table A.9: Convergence of the exact combinatorial score to $g(\rho)$. The column $N \times \varepsilon$ confirms the $O(1/N)$ error rate.

ρ	N	K	x_{comb} (exact)	$g(\rho)$	$N \times \varepsilon$
2	50	25	-1.4572	-1.3665	-4.54
2	100	50	-1.4118	-1.3665	-4.53
2	200	100	-1.3891	-1.3665	-4.52
2	500	250	-1.3755	-1.3665	-4.52
5	50	10	-5.2409	-4.9581	-14.14
5	100	20	-5.0953	-4.9581	-13.72
5	200	40	-5.0257	-4.9581	-13.52
5	500	100	-4.9849	-4.9581	-13.40

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Ashby, F. G., Alfonso-Reese, L. A., Turken, U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*. https://doi.org/10.1037/0033-295X.105.3.442

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual review of psychology*. https://doi.org/10.1146/ANNUREV.PSYCH.56.091103.070217

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–161. https://doi.org/10.1111/j.1749-6632.2010.05874.x

Gotelli, N. J., & Graves, G. R. (1996). *Null models in ecology*. Smithsonian Institution Press.

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). *Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13* (tech. rep. No. 97-02). University of Colorado, Boulder Institute of Cognitive Science. Boulder, CO.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*. https://arxiv.org/abs/1705.06950

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *262*(841), 23–81. https://doi.org/10.1098/rstb.1971.0078

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223. https://doi.org/10.1111/j.1551-6709.2011.01212.x

Navarro, D. J., & Kemp, C. (2017). None of the above: A bayesian account of the detection of novel categories. *Psychological Review*, *124*(5), 643–677. https://doi.org/10.1037/rev0000077

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, *4*(6), 661–682. https://doi.org/10.1002/hipo.450040605

Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026

Pitman, J. (2006). *Combinatorial stochastic processes* (Vol. 1875). Springer. https://doi.org/10.1007/b11601500

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*(3), 303–343. https://doi.org/10.1207/S15516709COG2603_6

Reiter, R. (1978). On closed world data bases. In H. Gallaire & J. Minker (Eds.), *Logic and data bases* (pp. 55–76). Plenum Press.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. https://doi.org/10.1037/a0020511

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., & Boult, T. E. (2013). Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1757–1772. https://doi.org/10.1109/TPAMI.2012.256

Shou, M. Z., Lei, S. W., Wang, W., Ghadiyaram, D., & Feiszli, M. (2021). Generic event boundary detection: A benchmark for event segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8075–8084. https://github.com/StanLei52/GEBD

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.*, *26*(3), 339–373. https://doi.org/10.1162/089120100561737

Teh, Y. W. (2006). A hierarchical bayesian language model based on pitman–yor processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. https://doi.org/10.1198/016214506000000302

Temme, N. M. (1993). Asymptotic estimates of stirling numbers. *Studies in Applied Mathematics*, *89*(3), 233–243. https://doi.org/10.1002/sapm1993893233

Tria, F., Loreto, V., & Servedio, V. D. P. (2018). Zipf's, heaps' and taylor's laws are determined by the expansion into the adjacent possible. *Entropy*, *20*(10), 752. https://doi.org/10.3390/e20100752

Tria, F., Loreto, V., Servedio, V. D. P., & Strogatz, S. H. (2014). The dynamics of correlated novelties. *Scientific Reports*, *4*, 5890. https://doi.org/10.1038/srep05890

Väänänen, J. (2007). *Dependence logic: A new approach to independence friendly logic*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611193

Wimsatt, W. C. (1987). False models as means to truer theories. In M. H. Nitecki & A. Hoffman (Eds.), *Neutral models in biology* (pp. 23–55). Oxford University Press.

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. https://doi.org/10.1016/j.tins.2011.06.006

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, *51*(3), 581–612. https://doi.org/10.1007/s10579-016-9343-x

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementsource.zip
- codeanddataarchive.zip