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.

Which level of abstraction, and when? Resource-rational granularity in inflection class learning

Maria Copot

m.copot@surrey.ac.uk

University of Surrey

Jayden Macklin-Cordes

University of Zurich

Research Article

Keywords: morphology, inflection classes, resource rationality, Bayesian nonparametric methods, category learning, granularity

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11362813/v1

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

Additional Declarations: No competing interests reported.

Which level of abstraction, and when? Resource-rational granularity in inflection class learning

Abstract

Morphological knowledge can be characterised at multiple levels of abstraction: the same facts about the lexicon can be stated as a single schema per inflectional macro-class, as intermediate sub-schemas over lexeme sets that share a given property, or as stored individual forms. This is usually cast as a choice of cognitive architecture: pure prototype, pure exemplar, or intermediate/multi-level systems. We reframe the question of the correct level of abstraction to model morphological knowledge as not one of architecture, but one where optimal level of abstraction is itself the variable of interest, to be derived from properties of the data. Under a resource-rational view (Anderson, 1991), humans build as much representational structure as the evidence justifies and no more, so the optimal granularity is inferred from the input, shifts as evidence accumulates, and may differ across regions of the lexicon. We treat the question of how we partition and represent lexical knowledge as a categorisation problem, and model it as Bayesian nonparametric inference in which granularity is a free parameter. We model trial-by-trial performance in an artificial-language experiment on plural allomorphy with exceptions with models that make different assumptions about granularity. On the basis of the results, we argue that granularity is thus productively thought of not a property of the morphological architecture to be discovered but a quantity inferred from evidence. The method speaks directly to theories of morphology which have an architecture where all levels of abstraction coexist, enabling them to make predictions about which grain a given speaker brings to bear on a particular linguistic task for a particular word.

Keywords: morphology, inflection classes, resource rationality, Bayesian nonparametric methods, category learning, granularity

1 Introduction

The same morphological facts can typically be captured by generalisations at different levels of abstraction. English past tense formation, for instance, can be described by a) a single productive rule that adds -ed to any verb stem, with all remaining forms listed as exceptions, b) by a set of intermediate schemas grouping irregular verbs into phonologically coherent clusters — string/strung, cling/clung, dig/dug share one

vowel change pattern, blow/blew, grow/grew share another — with -ed covering the remainder or c) by exhaustive item-by-item storage with no generalisations at all. All three descriptions are consistent with the observed data, they differ only in their granularity. A number of morphological theories have made this multi-level character of morphological knowledge central to their architecture. Construction Morphology (Booij, 2010), Relational Morphology (Jackendoff & Audring, 2019), and Network Morphology (Brown & Hippisley, 2012) all posit representations in which specific stored forms coexist with generalisations at varying levels of abstraction. The same insight, if less architecturally foregrounded, informs usage-based models of morphological organisation (Bybee, 2001, 2010) and probabilistic approaches to morphological productivity (Albright & Hayes, 2003). That multiple levels of generalisation coexist is, at this point, a well-established observation that has made its way into morphological theory. But it remains a descriptive observation, a characterisation of what morphological knowledge contains, not an account of how it emerges or how it is deployed. The data are compatible with many possible sets of granularities; existing theories do not specify which set a learner converges on, nor, once multiple levels of representation are available, which level is brought to bear in any given instance of producing or comprehending a novel form.

The Paradigm Cell Filling Problem (PCFP henceforth; Ackerman, Blevins, & Malouf, 2009) — how speakers use the inflected forms they know to produce and comprehend the ones they have not encountered — has been central to morphological research for decades. Before a speaker can perform the core task of the PCFP, a speaker must decide which existing words they should use as the basis for the task, in other words they need to decide on a word's analogical reference set. This decision is structurally a categorisation: a discrete assignment imposed on a continuous, multidimensional space of similarity between words, made from incomplete evidence. Framed this way (developed in Section 2.3), the question of which level of representation a speaker deploys becomes a question about grain: how finely the lexicon is partitioned into reference sets, and how that partition shifts as evidence accumulates.

Learning a category system under incomplete evidence is a problem of allocating representational resources: how many categories to maintain, and how much detail to encode within each, given limited and incrementally arriving data (Anderson, 1991). Applied to morphology, the trade-off is one of grain: a coarse representation that collapses an inflection class into a single summary captures its dominant pattern but cannot accommodate exceptions or subregularities, whereas a maximally fine-grained one that stores every form preserves all detail but makes each generalisation subject to the idiosyncrasies of whichever single item happens to lie closest to the novel word. The optimal resolution of this trade-off is not fixed: it depends on how much evidence the learner has accumulated and on the structure of the system being learned. Cognitive science offers a formal treatment of exactly this. Prototype and exemplar models are not competing architectures but limiting cases of a single clustering model - one cluster per class versus one per form - with intermediate grains in between, and the grain itself can be inferred from the data rather than fixed in advance (Griffiths, Canini, Sanborn, & Navarro, 2007, explained in Section 2.4). This reframing provides something that existing morphological theories lack: a mechanism for determining

which level of abstraction a learner should deploy, treating grain not as an architectural commitment but as an inference whose answer shifts as evidence accumulates.

In this paper, we import this framework into morphological theory. We model inflection class learning as Bayesian nonparametric categorisation, operationalising the question "which level of abstraction?" as inference over cluster granularity within a Hirearchical Dirichlet Process (henceforth, HDP). We compare three parameterisations of the model against human behaviour in an artificial language learning experiment: a prototype model that maintains a single cluster per inflection class, an exemplar model that stores every word form separately, and an intermediate model that infers its own clustering from the data. The experiment is designed to create divergent pressures for different levels of abstraction: the morphological system contains phonological prototypes that favour schema formation alongside exceptions that require item-specific storage. By fitting all three models to trial-by-trial responses across the course of learning, we can ask not only which granularity best accounts for human behaviour, but how the answer changes as learners accumulate evidence.

More broadly, this approach reframes a question that has occupied morphological theory for decades. The debate over whether morphological knowledge is rule-based or analogical, abstract or item-specific, has largely proceeded as a debate over cognitive architecture — as though the answer should be the same for all speakers, all systems, and all stages of learning. The framework we propose replaces this with a different question: what predicts optimal representational granularity? The answer, we argue, is not a fixed property of the cognitive system but a variable one, shaped by the quantity of evidence available to the learner, the distributional structure of the morphological system, and the interaction between the two. The present study takes a first step by isolating one of these factors — amount of evidence — in a controlled experimental setting, and demonstrating that human learners' representations are best captured by an intermediate granularity that is neither purely abstract nor purely item-specific, and is allowed to evolve as more evidence becomes available.

2 Background

2.1 The granularity question in morphological theory

Different theories of morphology have different goals. Some aim primarily to characterise the patterns attested across the world's languages, mapping the space of possible systems: for these theories, the granularity at which generalisations are formulated is a parameter of the analysis, set either axiomatically or downstream of the phenomena of interest. These theories contrast with those which take cognitive plausibility, or compatibility with theories of cognition, as a desideratum. For this second group, the question of granularity becomes substantive: at what level of abstraction are morphological generalisations represented in the cognitive system? The English past tense was the first substantial battleground for this question. At one pole, exemplar approaches (Bybee, 2001; Bybee & Moder, 1983; Bybee & Slobin, 1982) held that morphological knowledge consists of stored word forms, with schemas emerging as epiphenomena of the density and coherence of those exemplars. Regularity, in this view, is a property of how the exemplar cloud is distributed, not a separate level

of representation. At the opposite pole, Pinker and Prince (1988) proposed a qualitative architectural split: irregular forms stored individually in declarative memory, regular inflection computed by a single symbolic rule, later grounded neuroanatomically in the procedural memory system (Ullman, 2001, 2004). In granularity terms the two occupy opposite extremes: the dual-mechanism architecture combines the coarsest possible representation for regulars with something approaching the finest for irregulars, while the exemplar architecture lets coarser structure emerge from a maximally fine-grained storage. The points of tension between them brought into focus the kinds of data the field would eventually have to accommodate: partial regularities, gradient productivity, and phonologically coherent subpatterns within a broader regular class.[1] Parallel trajectories in adjacent literatures arrived at similar diagnoses: in word recognition, obligatory decomposition (Taft & Forster, 1975) and full listing (Butterworth, 1983) gave way to dual-route models in which granularity is variable and item-conditioned (Baayen, Dijkstra, & Schreuder, 1997; Caramazza, Laudanna, & Romani, 1988; Schreuder & Baayen, 1995); in the cognitive science of categorisation, the prototype-versus-exemplar opposition was reframed by the demonstration that the two are limiting cases of a single Bayesian clustering model, a result we return to in §2.4.

The first response to this residue, within morphological theory, was to build multiple levels of representation into the architecture itself, treating the coexistence of item-specific and schematic information as a positive design choice rather than a problem to be argued away. Network Morphology (Brown & Hippisley, 2012) represents morphological knowledge as a default-inheritance network in which information at varying levels of specificity coexists, with more specific nodes overriding more general ones when both apply. Construction Morphology (Booij, 2010) treats morphological structure as a constructicon of form–meaning pairings at all degrees of abstraction, with strong cognitive and usage-based commitments developed across the broader Construction Grammar tradition (Goldberg, 2006, 2019), though whether information is in fact stored at multiple levels or only at the most abstract one remains a longstanding internal debate (Hilpert, 2021). Relational Morphology (Jackendoff & Audring, 2019) takes the most explicit position on this storage question, treating multi-level storage as a positive claim about the lexicon, with item-specific and schematic representations coexisting by design. The same insight, if less architecturally foregrounded, informs usage-based models of morphological organisation (Bybee, 2001, 2010) and probabilistic approaches to morphological productivity (Albright & Hayes, 2003). That multiple levels of generalisation coexist is, at this point, a well-established observation, and a substantial portion of recent morphological theorising has been shaped by accepting it.

These frameworks accommodate the existence of representations at multiple grain sizes, but they do not, in general, account for how a learner builds those representations from exposure to data, nor for which level a learner brings to bear in any given act of producing or comprehending a novel form. In granting the architecture

[1] Connectionist models (Rumelhart, McClelland, & PDP Research Group, 1986), sometimes grouped with exemplar approaches in retrospective summaries of the debate, dispensed with discrete stored units altogether in favour of distributed representations over subsymbolic features, and are therefore not characterisable in terms of granularity as discussed here, given their qualitatively different nature.

the capacity to support multiple levels, they leave granularity itself as a fixed property of that architecture, stipulated in advance, rather than as something inferred from, and responsive to, the evidence a learner has encountered. Consider a learner who maintains, say, three or four sub-clusters within the regular class, corresponding to phonologically coherent sub-patterns, while still treating these as variants of a broader generalisation: this is precisely the representational state that the polar positions could not naturally express, and that multi-level architectures permit. Once the architecture is granted to support representations at multiple grain sizes, then, a new question opens up: which level do learners converge on, for which subsets of the lexicon, on the basis of what evidence, and how does the answer shift as that evidence accumulates?

2.2 Toward granularity as a variable

Several lines of work within morphology have moved towards the question the last paragraph ends on, each tackling different parts of it, though never all at once. A first set of strands frame granularity as a quantity inferred from the data, rather than fixed in advance by the architecture. Once we think of grain as being inferred from the data in this way, the question of how our representations partition the lexicon becomes inextricably tied to the amount of evidence a learner has, and the cases it covers. Approaches to representations that infer quantities from the data are inherently theories of learning, even though the object they model is the static endpoint of the learning process, using an idealised learner. This highlights the importance of a separate line of work that makes the process of learning itself the object to be explained, where the trajectory of the inferred grain becomes the variable to be explained. What is missing is an account that combines the two for inflection class representation, which we aim to build in the sections that follow.

The first step in the direction of inferring grain from data has been in the form of a growing interest in computational models that can infer multiple, gradient rules, mirroring the putative structure of human morphological knowledge. The first in this line is Albright and Hayes (2003), who propose, as an explicit alternative to both the dual-mechanism model and pure analogy, a Minimal Generalization Learner that induces stochastic rules at varying degrees of specificity, each with a confidence score reflecting its reliability in the lexicon. Their wug-test data show that participant ratings of novel forms are sensitive to phonological shape even for regulars, in a way that a single-default rule cannot capture and a pure exemplar model cannot motivate. Kapatsinski (2018) pursues a structurally similar agenda via conditional inference trees, extracting general-to-specific schemas from the data, with the additional argument that learners are product-oriented, generalising over output patterns, rather than source-oriented, formulating input-to-output rules in the sense of Albright and Hayes. Both lines of work produce a multi-grain inventory of generalisations induced from data, though neither treats granularity itself as a single tuneable quantity. Hay and Baayen (2005) make a complementary point from the empirical side: gradient productivity, gradient decomposability, and gradient paradigmatic strength are pervasive features of morphological systems, and any adequate theory must accommodate them as continuous phenomena rather than as deviations from binary categories. Their review separates

the question of probabilistic versus deterministic structure from the question of symbolic versus subsymbolic implementation, arguing that gradience in the grammar is an empirical finding independent of any particular computational framework.

A different line of work supplies an explicit granularity parameter determining the grain at which representations are stored. Yang (2016) proposes the Tolerance Principle: a rule generalises if the number of exceptions is below $N/\ln(N)$, where N is the number of items in the relevant class, and otherwise items are stored individually. The principle addresses the granularity question directly, framing it as a binary choice between one cluster (a productive rule) and many (stored forms). The threshold itself is principled and falsifiable, and Schuler, Yang, and Newport (2016) provide experimental support for it in children's behaviour. The binary architecture of Yang's proposal conveys a belief that there is no gradience between the two granularities, fully stored or fully rule-based. Howver, in the same experiment by Schuler et al., the adult participants showed no such categorical contrast, instead matching the token frequency of the majority pattern (cf. Kam & Newport, 2005). Adult generalisation here is thus frequency-sensitive rather than all-or-nothing as would be predicted under the Tolerance principle, a graded departure from the binary prediction (though probability matching over input frequencies is a distinct phenomenon from the intermediate-granularity representations we turn to below). In a recent proposal closely related to our own, Dekker, Rasilo, and de Boer (2025) recast inflection-class learning as the unsupervised clustering of lexemes by phonological similarity, and apply Adaptive Resonance Theory (an incremental clustering algorithm in which each incoming form joins an existing category or, if none is close enough, creates a new one) to Latin, Portuguese, and Estonian verb forms, treating its vigilance parameter (the similarity threshold that decides between joining an existing cluster or creating a new one, and therefore how many classes emerge) as a granularity knob, and reporting best recovery of attested classes in a narrow band of vigilance. Their vigilance is set by the modeller rather than inferred from the learner's evidence, and their evaluation is against typological endpoints rather than the trajectory of an individual learner. We return in §2.3 to what changes when granularity is instead inferred from a learner's accumulating data and allowed to shift as learning unfolds. O'Donnell (2015)'s Fragment Grammars also follow this same thread: the model treats the storage-versus-generalisation tradeoff as Bayesian inference under a rational resource-allocation principle, using an Adaptor Grammar over a probabilistic context-free grammar where the inference target is which sub-computations to memoise (i.e. cache the result of frequently needed computations in memory) and which to compute on the fly. The model is fitted to corpus-level distributions, capturing the endpoint properties of an idealised optimal learner. O'Donnell's modelling approach is able to capture both the fairly sharp regular/irregular split in the English past tense, and the gradient productivity of English derivational morphology, demonstrating that a single Bayesian-nonparametric framework can produce the empirical signature of both gradient and categorical processes. The question O'Donnell addresses is one of productivity versus reuse for whole-form storage in derivational morphology – whether *coolness* is stored whole or computed from *cool* and *-ness* – and the unit of inference is the fragment. This is a different question, at a different level of the architecture, from how finely a learner partitions

an inflection class into analogically coherent sub-clusters, but it shares the underlying methodological commitment: storage and generalisation are the target of rational inference, not stipulated architectural primitives.

The strands discussed so far make granularity inferrable but characterise it as a property of a static system. A further line shifts the object of explanation of representation-oriented theories from the representation that is learned to the process of learning itself. This line treats learnability as a precondition on theoretical abstractions: an abstraction earns its place in the theory only if it can be shown to be learnable from usage.Divjak and Milin (2026) argue that the cornerstone notion of emergence in usage-based theories has been treated axiomatically rather than empirically substantiated, and propose that learning, understood as gradual change in knowledge, can serve as an operational definition of emergence: only what can be learned from exposure to usage can be said to have emerged from it, which constrains theory to abstractions demonstrably learnable from the input. Reframing the question from what must be learned to what can be learned shifts attention from the endpoint representation to the process by which it is acquired, and makes properties of that process (the shape of the learning curve, its sensitivity to input structure) candidate variables of interest in their own right for understanding representations.

Taken together, the pieces needed to treat granularity as a variable to be explained are already present in this literature, what is missing is an account that assembles them for the specific question of inflection class representations. The problem would benefit from a single framework in which granularity is treated as a continuous quantity, inferred from the data rather than fixed in advance, allowed to shift as evidence accumulates within an individual learner, and testable against the trajectory of that learner's behaviour. The subsections that follow assemble such a framework: first by recasting inflection class learning as a categorisation problem (§2.3), and then by importing Bayesian nonparametric methods that treat granularity itself as the object of inference (§2.4).

2.3 Inflection class learning as categorisation

Inflection class systems – groupings of lexemes that share inflectional behaviour – vary substantially across languages along several dimensions (Corbett, 2009; Stump & Finkel, 2013). Languages differ in the sheer number of classes they support, from the Romance verb system being traditionally characterised as having a two- or three-way split to the hundreds of paradigm types attested in the Oto-Manguean languages (Palancar, 2014) or in Finnic systems such as Estonian (Blevins, 2008). In spite of the extensive surface variation we can observe in inflectional systems, they have common properties that define the learner's task. First, classes are unevenly populated: a few large ones coexist with many small ones, and the cells of each lexeme's paradigm are themselves distributed in a zipfian way, so that most cells of most lexemes are encountered rarely or not at all (Blevins, Milin, & Ramscar, 2017). Second, systems are scaffolded by extensive paradigmatic implicative structure: the forms of a paradigm carry predictive information about one another (Ackerman & Malouf, 2013; Bonami & Beniamine, 2016). Together these mean that, whatever the language, a learner confronts the same two-part problem, always from a partial and skew evidence: we

must predict each lexeme's unseen forms from its seen ones, and we must do this by determining which other lexemes we know are relevant to reason from to help us perform the prediction.

A substantial tradition within morphological theory treats this mapping problem as fundamentally one of form generation. The Paradigm Cell Filling Problem (PCFP; Ackerman et al., 2009) frames the central question as inference from known forms to unknown ones within a paradigm: how does a speaker who knows some inflected forms of a lexeme produce, or comprehend, the forms they have not yet encountered? This is a reframing of older, classification-based approaches where inflection classes are primitives: a lexeme is sorted into one of a fixed inventory of classes and its unknown forms follow from the class it has been assigned to. The generation view turns inflection classes into epiphenomenal constructs: insofar as they are descriptively useful, they are a reification of recurrent patterns of paradigmatic implicative structure. The implicative structure programme (Beniamine, Bonami, & Luís, 2021; Blevins, 2016; Bonami & Beniamine, 2016) centers on quantifying the predictive relationships that hold between the cells of a paradigm and identifying which forms carry which inferential weight. The approach has yielded principled answers to questions that classification-oriented approaches have tended to leave implicit: that morphological systems are organised so as to keep average conditional entropy low, which begins to explain why class systems take the shapes they do (Ackerman & Malouf, 2013), and that the diachronic resilience of class systems, i.e. their resistance to wholesale analogical levelling into a single uniform class, can be derived from the PCFP itself, given a rational reasoner that attends to both associative and dissociative evidence (Round, Esher, & Beniamine, 2024). We take the generation view, with its implicative structure operationalisation, as the broader research programme within which the present work proceeds.

Framing inflection class assignment as a categorisation problem is not a return to the classification view that the PCFP frames itself against. It is, rather, a claim about what the generation programme presupposes: before a speaker can project from a lexeme's known forms to its unknown ones, they must first settle a logically prior question — which existing lexemes should form the basis for the task? Or, otherwise put, what is its analogical reference set? Filling a paradigm cell for an unfamiliar word requires having already identified the words whose behaviour licenses the inference, and that identification is, structurally, a categorisation: the assignment of a novel item to a set of existing items that share inflectional behaviour, from which the novel item's own inflectional properties are then predicted. Categorisation in this sense is prior to the analogical step of the PCFP. It does not displace implicative structure as an account of what class systems are or why they are shaped as they are, but supplies the reference set over which implicative inference operates. Whether that set is called a "class", an "implicative neighbourhood", a "reference set", or a "cluster" is a theoretical commitment that is independent of the operation itself.

The scope of this claim is deliberately narrow. We do not propose that inflection class learning is categorisation in any reductive sense, nor that the categorisation framing replaces the generation view as a theory of what inflection class systems are. We propose that one component of the problem – the assignment of novel items to

analogically coherent reference sets of some granularity – is tractable as categorisation, and that treating it this way brings the formal tools of the cognitive science of category learning to bear on the specific question raised in §2.1: how finely does a learner partition the lexical space into analogically coherent reference sets, and how does this partitioning shift as evidence accumulates? Broader questions about the nature of class systems remain untouched by this, and the tools we develop below could in principle be deployed within an implicative framing as readily as within a class-based one.

Analytical tools from cognitive categorisation are useful because of the particular structure of the task that is the PCFP. Speakers must navigate a continuous lexical space in which words are positioned by multidimensional similarity: phonological shape, lexical semantics, and the behaviour of derivationally related words, among other factors, all contribute to how close or distant two lexemes are. The immediate outcome of a PCFP-type inference, however, is discrete: in each instance just one of the available inflectional forms is selected for generation. The learner's task is thus to impose discrete outcomes on a continuous similarity space, using probabilistic and overlapping cues, with graded rather than sharp boundaries, and on the basis of necessarily incomplete evidence, since no speaker has encountered every word in every inflected form. These are precisely the defining properties of the categorisation problems studied in cognitive psychology (Ashby & Maddox, 2005; Nosofsky, 1986), whose formal apparatus can then productively carry over to the problem at hand.

Humans demonstrably acquire systems with this structure, and artificial-language learning research has mapped the conditions under which they do. Adult learners can acquire arbitrary class systems when phonological markers cue class membership on at least a subset of items, and learning is facilitated by cue salience, frequency, and word-initial position (Brooks, Braine, Catalano, Brody, & Sudhalter, 1993; Frigo & McDonald, 1998). When phonological and semantic cues compete, the weight adults assign to each depends on its salience and on which cue becomes available first during learning, earlier-available cues being prioritised even when less salient (Culbertson, Gagliardi, & Smith, 2017). What matters most for the present purpose, though, is that the generalisations made by learners, including their grain, change as evidence accumulates. When a dominant pattern coexists with exceptions, how readily learners extend it to novel items, rather than restricting it around its exceptions, depends on the naturalness of the pattern, the proportion of exceptions, and the amount of training the learner undergoes (Baer-Henney, Kügler, & van de Vijver, 2015). These same patterns are visible in naturalistic language acquisition, with children overextending a dominant inflectional pattern to items that are in fact exceptions, mastering the exceptions only later (Bybee & Slobin, 1982; Marcus et al., 1992). Across both naturalistic and artificial language learning, we see the same pattern: the balance a learner strikes between a broad, coarse-grained generalisation and item-specific treatment of its exceptions is set by the structure of the input and the evidence accumulated so far, with the broad generalisation in place before the exceptions detach from it. This can be thought of as a shift in grain over the course of learning – precisely the phenomenon a resource-rational account treats as inferred from accumulating evidence rather than fixed in advance.

These findings have been modelled within rational-inference frameworks, notably a hierarchical Bayesian model of the closely related verb-argument-construction problem, in which the learner balances a class-general tendency against item-specific behaviour according to the distributional evidence, and which fits both adult artificial-language data and children's developmental trajectories (Perfors, Tenenbaum, & Wonnacott, 2010; Wonnacott, Newport, & Tanenhaus, 2008). Such treatments, however, along with the class-discovery models discussed in §2.2, have concentrated on endpoint behaviour or developmental milestones – aggregate accuracy after training, qualitative match to attested inventories – rather than on the trajectory itself, taken as the object to be explained.

2.4 Bayesian nonparametric categorisation

The cognitive science of category learning supplies a framework in which the properties identified as missing at the end of §2.2 are available together. Its starting point is a long-standing observation (Anderson, 1991): categorisation can be productively analysed as rational resource allocation. A learner encountering a stream of observations must decide how many categories to maintain and how much detail to encode within each, given limited and incrementally arriving data, and constraints on memory and attention. The optimal allocation depends on the structure of the data: when a distinction is evidentially supported the learner is justified in maintaining it, and when it is not the learner is justified in collapsing it. Granularity, in this framing, is itself the resource being allocated: maintaining too many distinctions is costly when evidence is thin, while maintaining too few sacrifices information that is supported by evidence.

This framing has an important consequence for the long-running prototype-versus-exemplar debate in categorisation. Prototype models represent each category by a single summary computed across all its observed members (Reed, 1972; Rosch, 1975); exemplar models retain every encountered instance as a separate representational unit (Medin & Schaffer, 1978; Nosofsky, 1986). Anderson (1991), and subsequently Griffiths et al. (2007), showed that these are not competing cognitive architectures but limiting cases of a single Bayesian clustering model, differing only in how many clusters the learner posits per category. A prototype model emerges when the learner maintains one cluster per category; an exemplar model emerges when every observed item is assigned its own cluster - and intermediate configurations, with a few sub-clusters per category corresponding to internally coherent subgroups, arise naturally when the data contain sub-regularities that are neither fully general nor fully idiosyncratic. In this framework, the number of clusters per category is governed by a single concentration parameter, conventionally written α. Critically, α does not need to be fixed in advance: it can be inferred from the data, so that the optimal granularity is determined by what the evidence supports rather than stipulated by the modeller.

Griffiths et al. (2007) formalised this within the Hierarchical Dirichlet Process (HDP; Teh, Jordan, Beal, & Blei, 2006), a Bayesian nonparametric prior over clusterings – nonparametric in the technical sense that the number of clusters is not bounded in advance but grows with the data as the evidence warrants. Its hierarchical structure is well suited to the problem set out in §2.3. At the top level, observations are

partitioned into categories, and within each category they are further partitioned into clusters of varying coherence. Two concentration parameters govern the structure: α controls within-category clustering, low values pushing the model toward one cluster per category (prototype-like) and high values toward one cluster per item (exemplar-like); a second parameter, γ, governs whether clusters are shared across categories or kept distinct, a question we return to in §5. Both can be sampled from their posteriors rather than fixed, so that the model performs inference simultaneously over the clustering and over the granularity at which it is described. nonparametric Bayesian models of this kind have been applied to linguistic data before: to unsupervised word segmentation (Goldwater, Griffiths, & Johnson, 2009), to the discovery of inflectional paradigms from raw text (Dreyer & Eisner, 2011), and, as discussed in §2.2, to the storage-versus-computation tradeoff in the lexicon (O'Donnell, 2015), but with the aim of recovering structure from corpus-level distributions rather than tracking the granularity of an individual learner's representations as evidence accumulates, which is the target here.

Framed this way, the prototype, exemplar, and intermediate positions correspond to three distinct, testable settings of α. A learner whose representations are best described by a prototype model corresponds to a small value of α that does not change with evidence; one best described by an exemplar model corresponds to a large value that likewise does not change; and one best described by an intermediate/multi-level model corresponds to a value of α that is inferred from the data and permitted to shift as evidence accumulates. Models implementing these three parameterisations make distinguishable predictions not only about endpoint behaviour but about the shape of the learning trajectory: the two fixed-granularity models predict that the best-fitting account of a learner is stable across the course of learning, whereas the adaptive model predicts that the best-fitting granularity changes with exposure, in the direction the data warrant. §3 presents an artificial language learning experiment designed to yield trajectory data of the kind these predictions require, and §4 fits the three parameterisations to that data and reports the result.

3 Methodology

We perform a simple artificial language experiment to get data on the learning trajectory of morphological patterns that contain exceptions. We then model trial-by-trial data with multiple computational models with different assumptions about the granularity of speakers' mental representations, and compare their performance at different points in the learning process.

3.1 Experiment

The experiment asked participants to learn the rules of plural formation in a constructed language. The stimuli were constructed so that for each of the two possible plural patterns, all items except for one matched a simple schema, with the one exceptional item fitting the schema for the opposite pattern. In line with previous experiments on learning generalisations and exceptions, we expect participants to first pick up on the two broad generalisations that govern the choice of plural allomorph,

Table 1 Stimuli used in the experiment

prototype	position modified	stimulus	exception
GYVUCO	-	gyvuco	
GYVUCO	1	zyvuco	
GYVUCO	2	gavuco	x
GYVUCO	3	gybuco	
GYVUCO	4	gyveco	
GYVUCO	5	gyvuto	
GYVUCO	6	gyvuci	
ZABETI	-	zabeti	
ZABETI	1	gabeti	
ZABETI	2	zybeti	
ZABETI	3	zaveti	
ZABETI	4	zabuti	
ZABETI	5	zabeci	x
ZABETI	6	zabeto	

overextending the generalisation to all items that match it, and eventually, through feedback, pick up on the fact that there is one exception to each pattern.

3.1.1 Items

Stimuli for the experiment are shown in Table 1. We constructed two prototypes of six letters each. To construct the stimuli, we take each prototype and we modify one letter at a time, flipping it to that characterising the position in the opposite prototype: in addition to the prototype string itself, the first stimulus belonging to the prototype GYVUCO is obtained by taking the first letter <g> and swapping it for the letter in first position of the other prototype, <z>, yielding *zyvuco*. This is repeated for each of the six positions. Finally, the seventh member of each prototype is the prototype string itself.

The prototype GYVUCO was associated with the suffix *-ma* in the plural, and the prototype ZABETI with the suffix *-re*. One stimulus within each prototype is exceptional: *gavuco* takes *-re* in the plural, and *zabeci* takes *-ma*. These same stimuli, with the same exceptions, were presented to all participants. From an experimental perspective, randomising the exceptions and the suffix associated with each prototype would have been ideal, but due to features of the modeling outlined in 3.2, these features needed to be fixed between participants. Pilot versions of the experiment were performed varying both the suffix associated with each prototype, and which stimulus was exceptional. The learning trajectory observed for these counterbalanced versions of the experiment appears to be largely independent of the particular suffix associated with each prototype and the choice of exceptional stimulus.

3.1.2 Procedure

A sample trial from the experiment is shown in Figure 1. Each trial consisted of showing participants in the matrix sentence *I saw one X*, signaling that the form is singular, followed by the sentence *I saw two ...*, prompting a forced choice between

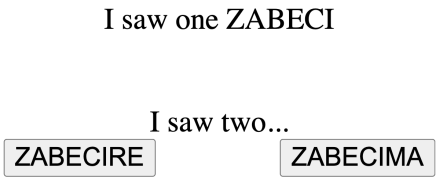


Fig. 1 A trial from the experiment

two possible plural forms. If the correct form was selected, the text "*Correct* ✓" would appear for three seconds and the correct button would become green, before automatically starting the following trial. If the incorrect form was selected, the text "*Incorrect* ×. *I saw two [correct_X.PL]. Try again.*" would appear, linger for two seconds, the correct button would become green and the incorrect one red, then both the colour and the text would disappear and prompt the participant to repeat the selection. Two incorrect selections on the same trial resulted in the text *"Try again"* appearing under the choices, freezing the selection for two seconds, then enabling it again until the correct selection was made.

The experiment consisted of 20 blocks, each containing 14 trials (one for each stimulus above) in a randomised order. So participants would see each stimulus 20 times over the course of the experiment.

Participants got two short breaks during the experiment, which they could skip if they preferred.

At the end of the experiment, participants were asked for demographic information (age, whether English is their only native language), and asked to complete a feedback form (how hard had they found the experiment? What did they think we were researching? What did they think the pattern was? Any additional comments?)

3.1.3 Participants

We recruited 43 participants from Prolific.co. All participants reported being over the age of 18, to have English as their mother tongue and to be living in the UK. Only one person reported fluency in non-English languages, but declined to report which. No participant was colourblind.

Time taken to complete the experiment ranged from 35 minutes to 1 hour and 40 minutes. The median time was 41 minutes. Participants were paid 12£ per hour.

3.2 Modeling

We fit three computational models to participants' trial-by-trial responses, each corresponding to a different assumption about the granularity of the mental representations underlying inflection class assignment. All three models are parametrisations of the category learning framework proposed by Griffiths et al. (2007), based on the Hierarchical Dirichlet Process (HDP; Teh et al. 2006), differing only in how many clusters they posit within each inflection class. By comparing the fit of these models to human

behaviour across successive stages of learning, we can assess which granularity of representation best accounts for participant performance, and how this changes as evidence accumulates. We additionally fit a fourth model that relaxes this assumption, allowing clusters to be shared across classes. Because it varies an orthogonal dimension – cross-class sharing rather than within-class granularity – rather than sitting on the granularity continuum, and because it proves near-indistinguishable from the adaptive model in fit (§4.5), we use it only to draw out lessons about cluster structure, returning in the discussion to the broader family of representations it makes available.

3.2.1 Feature representation

Each stimulus is represented as a vector of six binary features, one per letter position. For each position, the feature takes the value 0 if the letter matches the corresponding position in the prototype *gyvuco*, and 1 if it matches the prototype *zabeti*. Under this encoding, the prototype *gyvuco* is represented as $(0, 0, 0, 0, 0, 0)$ and *zabeti* as $(1, 1, 1, 1, 1, 1)$. Modified stimuli have one feature flipped: for instance, *zyvuco* (the first letter swapped from *g* to *z*) is encoded as $(1, 0, 0, 0, 0, 0)$. This captures the graded similarity structure of the stimuli – items sharing more features with a prototype are more similar to it – without building in explicit knowledge of which items are prototypical and which are exceptions.

3.2.2 The models

All models address the same prediction problem: given the labelled stimuli observed so far, how likely is a given word to take the suffix *-re* versus *-ma*? They differ in how stimuli within each suffix category are grouped into clusters.

In the first three models below, we fit $\gamma \to \infty$, so clusters are never shared between the two suffix classes. Each class maintains its own clusters and the models differ only in α, isolating granularity as the sole axis of comparison. Whether cross-class sharing is warranted at all is a separate question, governed by γ rather than α; we address it with a fourth, supplementary model, report the outcome in 4, and we return to the theoretical meaning of shared clusters in §5.)

Prototype model ***($\alpha \to 0$).***

All stimuli belonging to a given suffix class are assigned to a single cluster. This means each inflection class is represented by its central tendency: the model computes, for each class, the average feature values across all observed members, and compares a new stimulus to this average. The cluster assignments are deterministic and fixed.

Exemplar model ***($\alpha \to \infty$).***

Each observed stimulus token is assigned to its own cluster. The model retains every individual presentation of every word form as a distinct representation and categorises new stimuli by comparing them to all stored instances of each suffix class. Repeated presentations of the same word form contribute separate clusters rather than being merged, consistent with the formal equivalence between the exemplar model and kernel

density estimation (Ashby & Alfonso-Reese, 1995). As with the prototype model, the cluster assignments are deterministic and fixed.

Adaptive model (α sampled).

The concentration parameter α is treated as a free parameter estimated from the data. This allows the model to discover an intermediate number of clusters per category – potentially more than one (unlike the prototype model) but fewer than the total number of distinct stimuli (unlike the exemplar model). Because the cluster structure is not fixed in advance, inference over possible clusterings is performed using Markov chain Monte Carlo sampling (described below).

Shared model (α sampled, γ sampled).

The three models so far treat the two suffix classes as separate problems: each suffix class discovers its own internal groupings, and no grouping is ever shared between them. A learner might instead notice that a word patterns with items in the other class, e.g. that an exception resembles the phonological shape of words taking the opposite suffix, and represent that resemblance directly, rather than treating the exception as isolated. This is precisely what this fourth model allows. It is governed by a second parameter, γ, which controls whether groupings may be reused across suffix classes: the three models above hold $\gamma \to \infty$, forbidding any sharing, whereas the shared model infers γ from the data, letting a single cluster be drawn on by both classes when the evidence warrants it.

Formally, the shared model is identical to the adaptive model except that γ is also sampled (with a Gamma$(1, 1)$ prior); When a cluster is shared across the two classes, two things about it are kept separate. Its *shape*, the characteristic phonological profile that defines the cluster, i.e. how likely each stimulus feature is to take each value, is estimated from all of the cluster's members, pooled across both suffix classes. How much the cluster counts toward predicting a given class, however, depends only on how many of that suffix class's words fall in it. A shared cluster thus has a single shape, built from every word in it, while each suffix class keeps its own tally of how strongly it draws on that cluster. For instance, the exception *gavuco* (which takes $-re$) can join the same cluster as its *gyvuco*-shaped neighbours (which take $-ma$): the cluster's shape is then defined by all of them together, but it contributes to the $-re$ class only through *gavuco* and to the $-ma$ class through the *gyvuco*-family words. Because it varies γ on top of α, the shared model is not a further point on the granularity continuum, but a qualitatively different kind of representation. We therefore treat it not as a competitor to the three granularity models but as a test of whether the $\gamma \to \infty$ assumption distorts our conclusions, and as a probe of what, if anything, the two suffix classes share.

3.2.3 Computing category probabilities

All models are set to answer the same question: given the words the model has already sorted into clusters, and the suffix class each cluster belongs to, how likely is a novel stimulus $\mathbf{x}$ to take each suffix? This is done via a weighed 'vote': each suffix class votes for the new word through the clusters it contains, and a cluster counts for more when

it is large and when the new word closely resembles it. Formally, within each suffix class j we sum over the clusters k associated with that class, weighting each cluster's contribution by its relative size:

$$P(\mathbf{x} \mid c = j) = \sum_{k \in \mathcal{K}_j} \frac{M_{j,k}}{N_j} \cdot P(\mathbf{x} \mid z = k) \tag{1}$$

where $\mathcal{K}_j$ is the set of clusters within suffix class j, $M_{j,k}$ is the number of stimuli assigned to cluster k in suffix class j, and N_j is the total number of stimuli in suffix class j. The ratio $M_{j,k}/N_j$ is therefore the proportion of class j's stimuli that fall in cluster k, the share of the class that this cluster accounts for. The term $P(\mathbf{x} \mid z = k)$ is the probability of the stimulus under cluster k: here z is the cluster-assignment variable, with $z = k$ meaning membership in cluster k, so $P(\mathbf{x} \mid z = k)$ measures how well $\mathbf{x}$ fits that cluster's profile (defined in Equations 3–4 below). Each cluster thus contributes its resemblance to the new word, scaled by how much of the suffix class it represents, so that a large, well-populated cluster pulls the class's prediction toward it more strongly than a small one; summing these contributions over all of the class's clusters gives the class's overall score for the word.

Equation 1 applies directly to the prototype and exemplar models, whose clusterings are fixed in advance rather than inferred, by design. The prototype model puts all of a suffix class's stimuli into a single cluster, so each class has exactly one cluster ($M_{j,k}/N_j = 1$) and a new word is compared to the class's average. The exemplar model does the opposite, giving every stored stimulus its own cluster, so a class of N_j stimuli has N_j clusters, each of size one ($M_{j,k}/N_j = 1/N_j$), and a new word is compared to every remembered instance in turn. Because these clusterings are fixed and known in advance, the class scores follow directly from Equation 1 by a single, exact calculation, with nothing to estimate. For the adaptive model the clustering is instead inferred, so the same calculation is repeated over many sampled clusterings and averaged; see §3.2.4.

The adaptive model adds one ingredient: it holds back some probability for the possibility that the new word belongs to a pattern the model has not yet formed a cluster for, with the amount reserved governed by α. Equation 1 is therefore extended to:

$$P(\mathbf{x} \mid c = j) = \sum_{k \in \mathcal{K}_j} \frac{M_{j,k}}{N_j + \alpha} \cdot P(\mathbf{x} \mid z = k) + \frac{\alpha}{N_j + \alpha} \cdot P(\mathbf{x} \mid z = k^*) \tag{2}$$

This has the same basic shape as Equation 1, a sum over the suffix class's existing clusters, each contributing its resemblance $P(\mathbf{x} \mid z = k)$ to the new word, with two changes. First, the weights now divide by $N_j + \alpha$ rather than just N_j: adding α to the denominator sets aside a slice of the total probability for a cluster the model has not yet formed, so each existing cluster's share $M_{j,k}/(N_j + \alpha)$ is scaled down slightly to make room for it. Second, the extra term $\frac{\alpha}{N_j+\alpha} \cdot P(\mathbf{x} \mid z = k^*)$ governs what happens for that reserved slice: k^* denotes a brand-new cluster, and $\alpha/(N_j + \alpha)$ is the probability the model assigns to the new word starting a cluster of its own rather than joining an existing one. Because a new cluster has no members yet, its prediction $P(\mathbf{x} \mid z = k^*)$ falls back on the prior – every feature equally likely to be 0 or 1 – which

is $(1/2)^6$ (Equation 3 with $M_k = 0$ under the symmetric prior); a new cluster is thus maximally agnostic about what the word looks like.

The parameter α is what governs the granularity. When α is small the reserved slice of probability for unencountered custers is also small, so the model almost always assigns a new word to an existing cluster, yielding few, coarse clusters (prototype-like); when α is large the model readily opens new clusters, yielding many fine ones (exemplar-like). The adaptive model infers α from the data (§3.2.4), so granularity is set by the evidence rather than fixed in advance. For the prototype and exemplar models this new-cluster term drops out: as $\alpha \to 0$ and $\alpha \to \infty$ the probability of opening a new cluster goes to 0 and 1 respectively, so the partition is fixed (one cluster per class for the prototype model, or one per stimulus for the exemplar model) and we construct it directly rather than by sampling.[2]

How much a word resembles a cluster is judged feature by feature. For each of the six letter positions we ask how often the cluster's members took each value, so a cluster made up of words shaped like *gyvuco* places most of its probability on the *gyvuco* value at each position. Treating the six features as independent Bernoulli (i.e. two-valued) variables and integrating over the unknown probability with a symmetric Beta(μ, μ) prior gives, for feature dimension d:

$$P(x_d = v \mid z = k) = \frac{M_{k,v}^{(d)} + \mu}{M_k + 2\mu} \tag{3}$$

where $M_{k,v}^{(d)}$ is the number of stimuli in cluster k with value v on dimension d, and M_k is the total number of stimuli in cluster k. Setting the μ terms aside for a moment, this is just the proportion of the cluster's members that show value v at position d: if eight of a cluster's ten members have value 1 there, the cluster assigns probability 0.8 to a new word having value 1 at that position. The μ terms add a small *pseudo-count* to each of the two possible values – hence 2μ in the denominator, one μ per value – which biases the estimate toward 0.5 and, crucially and conversely, biases it away from 0 and 1: without it, a value the cluster happened never to show would be assigned probability 0, declaring any word with that value categorically incompatible with the cluster on the strength of a few observations. The size of μ controls how much smoothing is applied: a small μ lets the estimate track the observed frequencies closely, a large μ biases it towards the symmetric prior, near 0.5. μ is fitted for each participant (§3.2.7).

Equation 3 gives the probability for a single position. The probability of the whole word under the cluster is the product of these six per-position probabilities, since the

[2] We follow Griffiths et al. (2007) in identifying the prototype and exemplar models with these limits (building on Anderson 1991), fixing their clusterings rather than approximating the limit with a very small or very large finite α. A finite extreme value would only approximate it – a small α still occasionally splits a class, a large one occasionally merges stimuli – yielding a stochastic mixture near, but not equal to, the canonical models, with results depending on the arbitrary magnitude chosen and no gain but added computation. Fixing the clustering recovers those models exactly, matching Griffiths et al. (2007), for whom "the cluster structures for the prototype and exemplar models are fixed."

features are treated as independent:

$$P(\mathbf{x} \mid z = k) = \prod_{d=1}^{6} P(x_d \mid z = k) \tag{4}$$

This is the term $P(\mathbf{x} \mid z = k)$ that entered Equations 1 and 2 as the word's resemblance to cluster k. Because the six terms multiply, a word scores highly only if it matches the cluster across all positions: a poor match on even one position contributes a small factor that pulls down the similarity score.

Equations 1 and 2 give a score for each suffix class – how well the word fits that class – which is not yet a probability. The last step turns the two scores into probabilities over the two classes by normalising them: each class's score is divided by the sum of both, so that the two probabilities sum to one. This is Bayes' rule with equal prior probability on the two classes, so that the priors cancel and only the scores matter:

$$P(c = j \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c = j)}{\sum_{c'} P(\mathbf{x} \mid c = c')} \tag{5}$$

The result, $P(c = j \mid \mathbf{x})$, is the model's predicted probability that a given word takes suffix j – the quantity we compare against the participant's actual choice (after accounting for the guessing rate, §3.2.5).

The shared model computes the class probability $P(\mathbf{x} \mid c = j)$ with exactly the same expression as the adaptive model (Equation 2) – the same cluster weights, and the same reserved slice for a new cluster. What changes is how a cluster's feature profile is estimated. A cluster contributes in two separate ways to the probability that a word belongs to a suffix class: its shape – the feature profile in Equation 3, saying what a word in the cluster looks like – and its *weight* within a class, the mixing term $M_{j,k}/(N_j + \alpha)$ in Equation 2. In the adaptive model a cluster belongs to a single suffix class, so both are computed from that class' members alone. The shared model lets a cluster belong to both suffix classes at once, and estimates its shape by pooling its members across the two:

$$P(x_d = v \mid z = k) = \frac{M^{(d)}_{k,v} + \mu}{M_k + 2\mu} \qquad M_k = \sum_{j'} M_{j',k} \tag{6}$$

where the sum applies to both suffix classes j', so both M_k and the value count $M^{(d)}_{k,v}$ are now taken over the cluster's members in both classes rather than within one. A cluster shared between classes therefore has a single shape, informed by all of its members. Its weight, by contrast, is unchanged: the mixing term $M_{j,k}/(N_j + \alpha)$ still counts only the class-j members of the cluster, so each class decides how much to rely on the cluster on the basis of how many items in the cluster belong to the suffix class of interest. Concretely, the exception *gavuco* (which takes *-re*) can share a cluster with its *gyvuco*-shaped neighbours (which take *-ma*): the cluster's shape is then built from all of them together, but it counts toward the *-re* class only through *gavuco* and toward the *-ma* class through the *gyvuco*-family words. When no cluster

is shared across classes, $M_k = M_{j,k}$ and the shared model reduces exactly to the adaptive model.

3.2.4 Inference for the adaptive and shared models

For the prototype and exemplar models the clustering is fixed, so the class probabilities of Section 3.2.3 follow directly. The adaptive and shared models instead treat the clustering as unknown and infer it from the data: the number and composition of the clusters, and with them the concentration parameters α (and, for the shared model, γ), are not given but estimated. There is no closed form for the resulting average over possible clusterings, so we approximate it by drawing samples of plausible clusterings and averaging the class probabilities across them.

Because the number of possible clusterings is vast, the average over them in Equation 2 cannot be computed exactly. We approximate it by Markov chain Monte Carlo (MCMC): rather than enumerate clusterings, we generate a sequence of them in which each is a local revision of the previous one, constructed so that, over the sequence, clusterings are visited with frequency proportional to their posterior probability – that is, to how well they account for the data. A thinned subset of this sequence then serves as a sample from the posterior over clusterings, and the quantities of interest are averaged across it.

The sampler alternates two updates. In the first, the concentration parameters are held fixed and each stimulus is reassigned to a cluster in turn, following the Chinese Restaurant Process representation of the HDP (Teh et al., 2006): a stimulus joins an existing cluster with probability proportional to that cluster's current size, and forms a new one with probability governed by α; in the shared model, clusters may additionally be reused across the two suffix classes, to an extent governed by γ. In the second, the clustering is held fixed and the concentration parameters are themselves resampled – α, which controls the model's preference for few clusters over many, and, in the shared model, γ, which controls its readiness to reuse a cluster across classes – by the auxiliary-variable method of Teh et al. (2006). Both parameters are given weakly informative Gamma$(1, 1)$ priors, so that their values are determined by the data. Iterating the two updates allows the clustering and the concentration parameters to be inferred jointly.

Inference is carried out separately for each participant $\times$ segment. Because the chain is initialised from an arbitrary clustering, its early states are unrepresentative of the posterior. We therefore discard an initial burn-in and begin collecting samples only once the chain has approached its stationary distribution. Because successive states are strongly autocorrelated, we retain only one sample per fixed number of sweeps (thinning), so that the collected samples are approximately independent. The chain is advanced continuously rather than reinitialised between updates, so the burn-in is incurred only once. For the grid search over μ and g (Section 3.2.7) we use a shorter run of 12 iterations – each comprising 30 sweeps over the cluster assignments followed by one resampling of the concentration parameters – discarding the first 8 and retaining the remaining 4 samples. For the final predictions we use a longer run of 40 iterations of 40 sweeps, discarding the first 15 and retaining 25. Category probabilities are computed under each retained sample and averaged, so that the

prediction marginalises over the posterior uncertainty in the clustering rather than conditioning on a single partition.

3.2.5 Guessing parameter

To allow for the possibility that participants respond randomly on some proportion of trials, we include a guessing parameter g. On each trial, the probability of the participant's observed response is modelled as a mixture of the model's predicted probability and a uniform guess:

$$P(\text{response}) = (1 - g) \cdot P_{\text{model}}(\text{response}) + g \cdot 0.5 \tag{7}$$

3.2.6 Training and evaluation

We divide the 20 experimental blocks into 10 segments of two blocks each. For each segment s, the model is trained on all correctly labelled data from blocks 1 through $2s$ – that is, all trials the participant has encountered up to the end of that segment, with the correct suffix as the training label regardless of the participant's actual response. The model is then used to predict the participant's response on each trial in the final block of the segment (block $2s$). As s increases, the models have access to progressively more evidence, allowing us to track how model fit changes as learning unfolds. For the adaptive and shared models we additionally record, at each segment, the cluster membership of every stimulus in each retained posterior sample. Because cluster labels are not comparable across samples (label switching), we summarise these memberships by the posterior probability that each pair of stimuli is assigned to the same cluster, which we use in Section 4.4 to examine how the inferred grouping of stimuli evolves over the course of learning.

3.2.7 Parameter fitting

Three parameters are fitted for each participant separately: the symmetric Beta prior parameter μ, the guessing rate g, and the cluster granularity parameter α. The parameters μ and g are selected by grid search to maximise the total log-likelihood of the participant's observed responses across all 10 segments. The search grid is as follows:

- $\mu \in \{0.01, 0.1, 0.5, 1, 5, 20\}$
- $g \in \{0.01, 0.05, 0.1, 0.2, 0.3\}$

For the prototype and exemplar models, the clustering is fixed directly (one cluster per class; one cluster per token), and only μ and g are fitted by grid search. For the adaptive model, μ and g are fitted by grid search over the same ranges, but α is not grid-searched: instead, it is sampled from its posterior on each outer Gibbs iteration, jointly with the cluster assignments, and marginalised over when computing category probabilities. This means the adaptive model has the same number of fitted free parameters as the prototype and exemplar models (μ and g), while α is treated as a latent variable inferred from the data rather than a fixed hyperparameter.

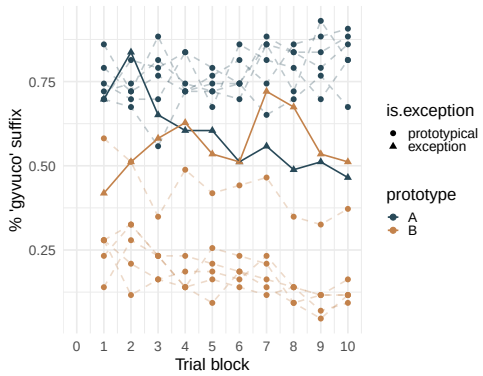


Fig. 2 Performance aggregated by stimulus for the full cohort of 43 participants

4 Results

4.1 Raw participant data

Figure 2 shows human performance aggregated by stimulus for the entire participant set, with stimuli marked for the prototype they belong to, and for whether they have exceptional inflectional behaviour. The y-axis represents the percentage of participants that chose the suffix associated with the *gyvuco* prototype in a given trial for each stimulus – if all humans gave the correct response on every trial, we would expect all prototypical stimuli associated with *gyvuco* and the *zabeti* exceptional stimulus to have value 1, and all prototypical stimuli associated with *zabeti* and the *gyvuco* exceptional stimulus to have value 0. What we see is a progressive polarisation of the prototypical stimuli, and a much slower inversion of the exceptions: the exceptions first pattern with their prototype, and progressively move towards more frequent responses for the correct, non-prototypical stimulus. However, while the trajectory of the prototypes broadly follows expectations, the trajectory of the exceptions appears very noisy. Additionally, there is prototype B prototypical stimulus that appears to get assigned the prototype A suffix more often than the rest: this is *zabeto*, whose non-canonicity is defined by its final segment being switched to that of prototype A. Equally, the prototype A prototypical suffix on which participants were least accurate is *gyvuci*, whose-noncanonicity is defined by its final segment being switched to that of the opposite prototype. It appears that the final segment of the singular is given particular weight when native English speakers are asked to perform linguistic classification tasks. This bias is not surprising, as when plural suffix allomorphy is phonologically conditioned in real languages, the conditioning environment is disproportionately likely to be stem-final.

It is possible that some of the participants may not have noticed or learned the designed patterns that govern plural allomorphy in these stimuli, and might have been holding other generalisations (such as the participants who identified the final segment as the determinant of category membership, as hinted above) or might have struggled to remember which words were the exceptions from one trial to the next. We wish to restrict the rest of the analysis to participants that appear to have latched onto

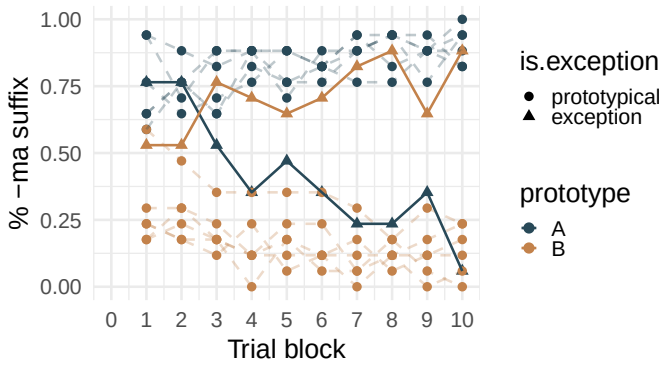


Fig. 3 Performance aggregated by stimulus for the subset of successful learners

the intended generalisations and exceptions intended for the data. We operationalise such participants as those who show improvement on the exceptions over the course of the experiment. To identify them, we fit a `brms` (Bürkner, 2021) regression to the exceptional stimuli only, with the formula $is.correct \sim 0 + (Trial \mid participant)$, where is.correct is a binary variable taking the value 1 when the participant's answer matches the intended answer for that stimulus, and Trial is centered. From the fitted model we take, for each participant, the posterior probability that their Trial slope is positive. Because participant inclusion is a screening decision rather than a hypothesis test, we do not require this probability to clear a conventional significance threshold; instead we retain participants for whom it exceeds 0.75, a value that falls in a natural gap in the distribution of these probabilities across participants. Of the 43 initial participants, 17 meet this criterion. A sensitivity analysis is conducted in the code accompanying this paper (available to reviewers in the attached files).

The response pattern for the subset of successful learners looks much cleaner than for the broader cohort. Figure 3 shows immediate differentiation for prototypical stimuli, and fairly early crossover for the exceptions, which by the end of the experiment receive their intended suffix with the same confidence as the prototypical stimuli.

To summarise accuracy over time on different stimulus types for this select set of participants, we fit a logistic `brms` (Bürkner, 2021) regression with the formula $is.correct \sim Trial * is.exception + (1 \mid participant)$. Trial is centered; is.exception is treatment coded, with the baseline set to 0 (the prototypical stimuli). The model output is shown in Table 2, and the conditional effects plot of the model in Figure 4.

The picture that emerges cleanly summarises the observed pattern: prototypical stimuli are on average much easier to get right than exceptions from the very start, and while accuracy on both improves over time (the Trial coefficient), the speed of improvement is faster for the exceptional stimuli (Trial:is.exception, for which all of the posterior is comfortably above 0). Early in the experiment, exceptional stimuli are more likely to receive the suffix corresponding to the prototype they match (the model predicts a probability below .5), but after participants have received enough feedback(Figure 4), they begin to match the exceptional stimuli with their intended suffix.

Table 2 Coefficients of a bayesian regression model predicting whether the participant has selected the intended suffix for the stimulus or not, by trial and stimulus type

	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	0.94	0.21	0.52	1.35
Trial	0.07	0.01	0.06	0.09
stimulus_type-exception	-1.68	0.20	-2.07	-1.29
Trial:stimulus_type-exception	0.06	0.02	0.03	0.10

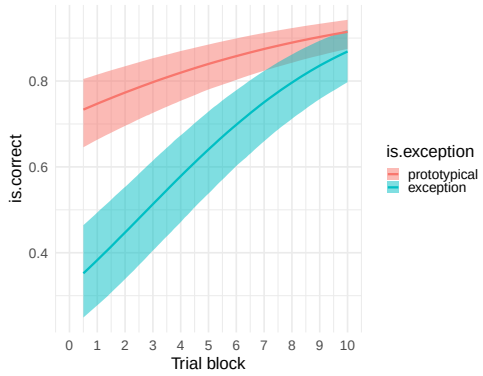


Fig. 4 Conditional effects plot of the `brms` model

This pattern will not be surprising to the reader, it matches both intuition and what has been observed in studies that track the learning of irregularity: the regular pattern is easy to learn due to its type frequency, and it gets overextended to irregular items until the learner acquires enough evidence for their surprising behaviour, after which performance improves as evidence accumulates.

4.2 Model fit and metrics

The three models make differing predictions about participant accuracy on prototypical and exceptional stimuli over the course of the experiment. Figure 5 shows the probability predicted by each model for each stimulus type over trials. These figures are the average predicted probability output by each model – the free parameters of the models have been set to what has been fit for each participant. All models suggest participants will start off strong on prototypical stimuli (adaptive and prototype model anticipate a small improvement before plateauing, exemplar accuracy rate does not change). All models predict that performance on the exceptions will start off weaker than on the prototypical stimuli. The exemplar model predicts the smallest difference between prototypical and exceptional stimuli, both with starting accuracies above chance, with accuracy remaining constant. The prototype model predicts that accuracy on exceptions will actually decrease as the experiment goes on, while the adaptive model predicts it will improve. It is on the exceptional stimuli that the three

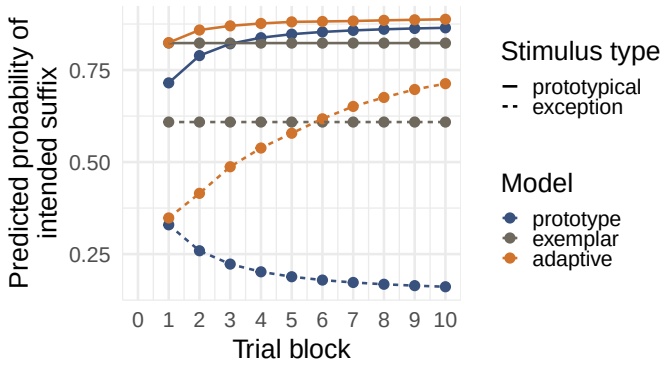


Fig. 5 Probability assigned by different models for participant accuracy on the intended suffix.

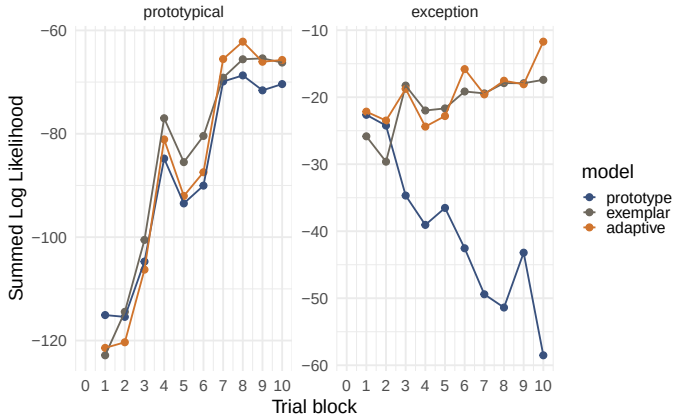


Fig. 6 Log likelihood by model, stimulus type, trial

models make their most divergent predictions, and so this is the data subset that we will focus on for adjudicating fit.

We begin by reporting log likelihoods of participant responses given each model, split by stimulus, as the number of trials grows, as this is the standard metric in computational cognitive science. A higher log likelihood means the data is better fit by the model. Figure 6 shows very close performance by the three models on prototypical stimuli. On the exceptions, the prototype model clearly stands out, assigning the exception to its prototype and making increasingly more confidently wrong predictions. The exemplar and adaptive model have comparable log likelihoods.

Before interpreting the closeness of the exemplar and adaptive models as evidence for their equivalence, it is important to examine a mathematical property of log-likelihood that bears directly on model comparison in this setting. Log-likelihood rewards confidence when the model is right and punishes it when the model is wrong, but asymmetrically, as visualised in Figure 7. The penalty for a confident wrong prediction grows without bound as the predicted probability approaches zero, while the

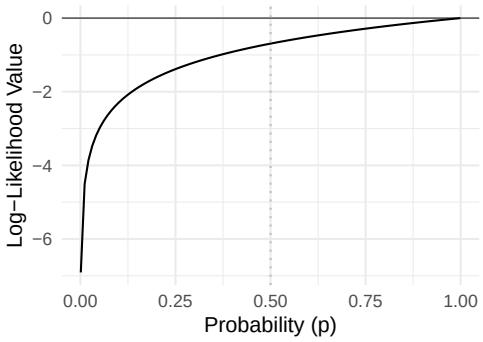


Fig. 7 How log likelihood varies given different input probabilities.

reward for a confident correct prediction is bounded (approaching zero as the probability approaches one). So what matters for performing well on log likelihood is not so much being right, or being well-calibrated, but never being confidently wrong. A model that hedges — assigning probabilities near 0.5 on every trial — is guaranteed never to incur catastrophic penalties, even though it never demonstrates learning. A model that commits to confident predictions is penalised severely when those predictions are wrong, even if its average accuracy is higher. In our data, this asymmetry matters because the prototype model and the adaptive model both make confident predictions — the prototype model confidently wrong on exceptions, the adaptive model first confidently wrong and then progressively confidently right. The exemplar model, by contrast, produces predictions that are much more consistently closer to 0.5, rarely committing strongly in either direction. Log-likelihood rewards this hedging: the exemplar is never catastrophically wrong on exceptions because it never strongly commits to a prediction. This has a consequence for intepretation: the exemplar model's competitiveness on log-likelihood is a property of the metric, not evidence that learners store exemplars. Its near-parity is is the result of predictions that do not deviate too far from 0.5 rather than by anticipating what participants do meaningfully better than competitors.

Because log-likelihood's asymmetry makes it a poor fit for this data, we ask whether the ranking of the models survives under measures that treat confidence differently. We consider the mean probability each model assigns to the participant's actual response, which is symmetric around 0.5 and simply rewards a model for anticipating what the participant did with greater confidence. We evaluate each both on the exceptional stimuli, on which the models make their most divergent predictions, and over all stimuli.

Figure 8 shows the mean probability assigned to the human's response, disaggregated by stimulus type. On the exceptions the adaptive model assigns consistently higher probability to the participant's response than either alternative across most of the experiment, while the prototype model, consistent with its anti-learning pattern, is worst and declines with training; on the prototypical stimuli the three models are comparable.

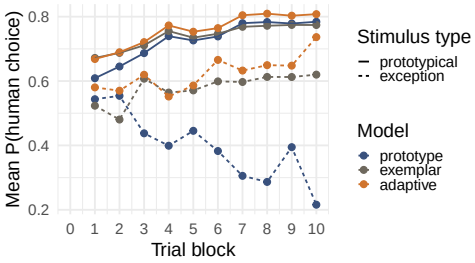


Fig. 8 Mean probability assigned by the model to the stimulus chosen by each participant on each trial block

Table 3 summarises how the models rank under each measure. On the exceptions, the measures agree that the adaptive model fits best, the exemplar model second, and the prototype model a distant third. Over the prototypical stimuli, all models perform comparably on log likelihood, and while they remain close on P(hum), the prototypical model is the worst performing, followed by the exemplar, with the adaptive model scoring slightly higher.

Table 3 Mean model fit by stimulus type, averaged over trials and successful learners: mean log-likelihood and mean probability assigned to the human's response (both higher is better). On the prototypical stimuli the three models are near-identical; on the exceptions the adaptive model leads on both measures

	Prototypical		Exception	
Model	LL	P(hum)	LL	P(hum)
prototype	−0.43	0.73	−1.18	0.40
exemplar	−0.42	0.74	−0.62	0.58
adaptive	−0.43	0.76	−0.57	0.62

The advantage of the adaptive model is, moreover, conservatively estimated. All three models were fitted per participant on the same two free parameters, μ and g, so the prototype and exemplar models can use smoothing and guessing to soften their predictions toward the human data, approximating the adaptive model's trajectory by parametric means rather than by adjusting granularity. That the adaptive model leads despite this latitude, and leads on P(hum), which unlike log-likelihood confident hedging cannot inflate, means the comparison likely understates rather than overstates the gap between an inferred granularity and a fixed one.

4.3 Patterns by participant

The section above showed overall model behaviour over segments. This section presents participant-level generalisations. Firstly, the overall behaviour of different

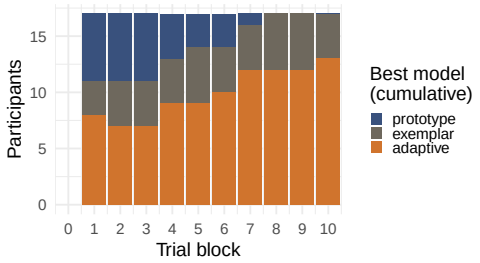


Fig. 9 The model that provides the best fit to participant behaviour cumulatively up to each trial block

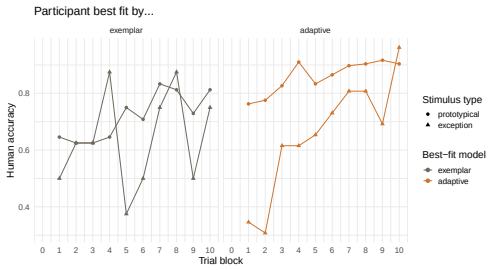


Fig. 10 Percentage of instances of each participant choosing the intended suffix for each stimulus, divided by stimulus type, and by which model best fit the participant's behaviour

participants was best captured by different models: on the basis of log likelihood, the adaptive model is the best fit for the large majority of learners (13), the exemplar model for a small number (4), and no participant is best fit by the prototype model. More tellingly, the best-fitting model is not constant over the course of the experiment (Figure 9): early on, a larger share of participants' responses are best matched by the prototype model, and this share diminishes as more data accumulates, with the best-fitting model shifting clearly towards the adaptive by the end, with a few participants who remain best fit by the exemplar model throughout.

Given that the adaptive model provides the best aggregate account of the exceptions (Section 4.2), it is worth asking why a minority of learners are nonetheless better fit by the exemplar model. Two possibilities present themselves: that these participants' representations really are better described by a different process, or that the difference in model fit is perhaps an artefact of a difference in the extent to which participants learned the intended pattern.

Figure 10 shows the percentage of stimuli for which each participant chose the intended suffix, split by stimulus type and by the model that best fits their behaviour. Everyone improves on the prototypical stimuli over the course of the experiment, though for the participants who are best fit by the exemplar model, accuracy remains overall lower. On the exceptions, however, the two groups diverge: participants best fit by the adaptive model show progressive improvement on the exceptions, which by the end of the experiment match the accuracy of the prototypical stimuli, whereas participants best fit by the exemplar model have somewhat lower accuracy on the

prototypical stimuli and an accuracy on the exceptions that hovers around chance and does not meaningfully improve. These appear to be participants whose responses are comparatively stochastic. The difference in model fit thus likely reflects a difference in the amount of learning, rather than a distinct underlying representation.

4.4 What clusters does the adaptive model infer?

Because the adaptive model infers its own clustering rather than being given one, it lets us observe not only how well it fits but what representation it settles on, and how that representation changes as evidence accumulates. Figure 11 shows the average number of clusters the adaptive model maintains at each stage of the experiment. It begins at just under 4 and rises to around 4.7 over the first two-thirds of the experiment before levelling off. The direction of this change is the central prediction of the framework: as evidence accumulates, the granularity the model infers increases, the learner coming to support more distinctions than the sparse early data would justify. If each learner converged on the cleanest representation the system admits, they would begin with two clusters (one per suffix class, each holding that class's prototypical items) and, on encountering the two exceptions, split off a cluster for each, plateauing at four. What we observe departs from this idealisation in both its level and its range: the count begins nearer four than two, and rises by about one cluster on average over the course of the experiment.

The departure is the result of the value being an average, and of how the clusters are fit. The measure plotted is first an average over the model's posterior uncertainty about the clustering for each learner, and over learners whose individual solutions differ. The number of clusters being higher than these idealised counts is the result of two factors. First, the regular items within a class are not identical; each is the class prototype with one feature altered, so the model can intermittently posit finer sub-structure among them, and the concentration parameter leaves some standing probability on an additional cluster even when the data do not clearly demand one. Second, the separation of the exceptions is partial and uneven across learners – not every learner fully and consistently splits both exceptions from their prototype – so their contribution to the average falls short of the idealised two extra clusters. The absolute number of clusters is thus a blunt summary, what is informative is the trajectory of the measure, which rises as evidence accumulates, and the composition of the clusters, which is discussed below.

To understand how the model forms clusters, we look at consensus co-assignment: for each pair of stimuli, the proportion of posterior samples in which they are grouped into the same cluster (Section 3.2.6). Figure 12 shows this co-assignment, averaged over learners, at different points in experiment. The prototypical stimuli of each suffix class remain strongly co-assigned throughout, forming a coherent group for each class from the outset. The exceptions, by contrast, begin grouped with the prototypical members of their own suffix class and progressively separate from them as the experiment proceeds.

We can quantify this separation directly. For each exception, Figure 13 tracks its mean co-assignment with the prototypical members of its own suffix class, alongside an isolation index (one minus its mean co-assignment with all other stimuli). Over

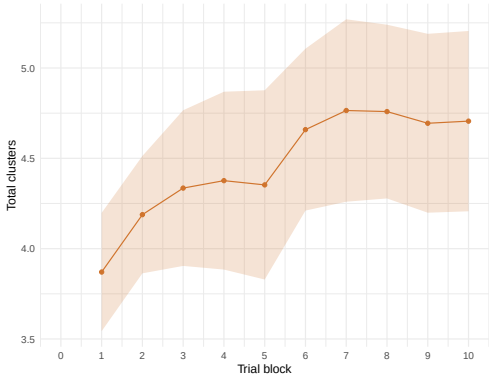


Fig. 11 Average number of clusters inferred by the adaptive model, by trial block (ribbon shows ±1 standard error)

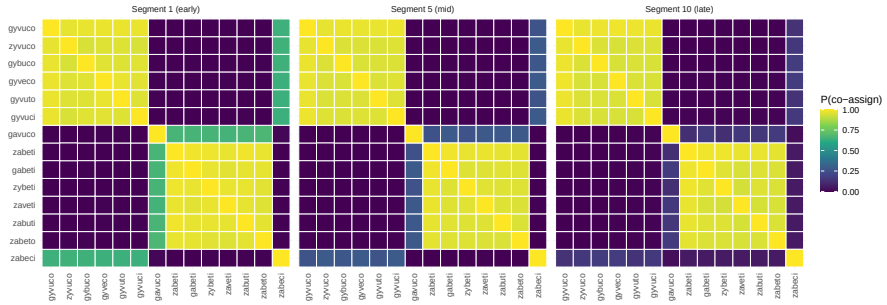


Fig. 12 Consensus co-assignment (proportion of posterior samples in which two stimuli share a cluster), averaged over successful learners, at an early, middle, and late segment. Exceptions are *gavuco* and *zabeci*.

the course of the experiment the first falls and the second rises: the exceptions detach from the words they initially clustered with and come to occupy clusters of their own. The separation is partial – even at the end the exceptions retain some co-assignment with their suffix-class neighbours – but its direction is clear and consistent across both exceptions.

4.5 The shared model

The three models compared so far keep the two suffix classes strictly separate. The shared model relaxes this by inferring γ, so that a cluster may be shared across the two classes (Section 3.2). Because it varies γ, it is not a further point on the granularity continuum; we treat it not as a competitor but as a check on the assumption that $\gamma \to \infty$ that is built into the other three models.

The shared model is indistinguishable from the adaptive model on fit to human responses: the per-participant difference in summed log-likelihood is close to zero, and the shared model assigns higher probability to the participant's response on only

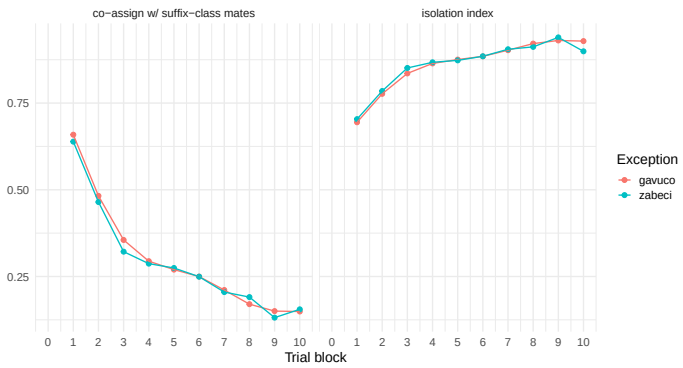


Fig. 13 For each exception, mean co-assignment with the prototypical members of its own suffix class, and an isolation index (1− mean co-assignment with all other stimuli), by trial block.

about half of participant – segment cells. Fixing $\gamma \to \infty$ in the primary comparison therefore doesn't result in a loss of accuracy. This near-equivalence is partly the result of structural factors: cluster sharing re-estimates only a cluster's shape, by pooling its members across classes, while the class-specific weight through which a cluster predicts a suffix is unchanged. Because each exception differs from the opposite class's prototype by a single feature, pooling it with its phonological neighbours does not really affect the overall shape for the cluster. The two models therefore issue almost the same suffix predictions here by construction, and the tie in performance should not be read as evidence against cross-class sharing in general. We return to this issue in Section 5.1.

What the shared model shows is that, when sharing is permitted, the expected cross-class structure emerges. The clusters that end up shared are those linking each exception to the prototypical members of the *opposite* suffix class – its phonological neighbours: *gavuco* co-clusters with the *gyvuco* family and *zabeci* with the *zabeti* family (mean cross-class co-assignment modest but consistent, around 0.18 for the strongest pairs), and the two exceptions cluster with one another. Given the freedom, that is, the model represents each exception partly by the phonological group it resembles, across the very class boundary its suffix defies. We return to what this structure might mean in Section 5.

5 Discussion

The paper shows that the level of abstraction that best describes a language user's representations is not fixed but shifts from coarse to fine as evidence accumulates. This shift is can be inferred both in the models' comparative fit and in the clustering the adaptive model infers, where the exceptions detach from the dominant pattern. As a result, only a model that is permitted to let its granularity vary can accurately capture the trajectory. These findings link to the long debate over whether morphological knowledge is rule-based or item-specific, abstract or analogical. The debate has largely been conducted as a dispute over architecture, as though there were a single correct grain at which the system stores and deploys its generalisations, the same for every

speaker, system, and stage of learning. We want to suggest that the question is posed at the wrong level. Granularity is not a fixed property of the cognitive architecture to be discovered but a quantity to be inferred, whose value depends on how much evidence a learner has and what structure that evidence carries. Reframed this way, "which level of abstraction?" becomes "what granularity does this data, at this amount of evidence, warrant?", a question with a determinate, measurable, and dynamic answer.

5.1 A mechanism for multi-level theories

The multi-level theories surveyed in Section 2.1 established the representational premise this paper starts from: that item-specific and schematic knowledge coexist, and that a theory of the lexicon must accommodate generalisations at more than one grain. What these theories characterise is the standing content of the lexicon (i.e. which schemas it holds, at which levels of abstraction) and not which of those coexisting representations a language user brings to bear on different language tasks. The framework we have developed is directed at exactly this gap.

What the framework contributes is twofold. First, it supplies a mechanism that turns granularity selection from a stipulation into an inference: the Hierarchical Dirichlet Process derives, for a given learner and a given quantity of evidence, which grain the data warrant, and predicts how that grain shifts as evidence accumulates. Which level of representation is used becomes in this framework consequences of the learner's evidence rather than a primitive of the analysis. Second, it supplies a measure. Because the adaptive model infers its clustering, the posterior co-assignment of stimuli (Figures 12, 13) is an instrument for measuring which grain is being inferred.

Both contributions so far concern selection among grains that a multi-level architecture already contains. The same machinery can also speak to how the architecture is built. The three models compared above keep the two suffix classes strictly separate, so that a cluster belongs to one suffix class alone. The shared model relaxes this by inferring a parameter γ that allows a cluster to be shared across classes. When sharing is permitted, the structure that emerges is interpretable. The clusters the model chooses to share are those linking each exception to the prototypical members of the opposite suffix class, its phonological neighbours. Given the ability to do so, the model represents each exception partly by the phonological group it resembles, partly by its suffix choice. An exception so represented is co-motivated: partly idiosyncratic, partly inheriting the structure of the prototype it resembles, rather than stored as a wholly separate unit. Structures of this kind, i.e. multiple inheritance/multimotivation, are precisely those that inheritance- and construction-based theories stipulate as the basis of their representations. A nonparametric model that samples γ offers a way to induce them from learners' data instead. We see this as a potential fruitful direction for future research than strictly as a result of the current work. The tie in fit is, as noted in Section 4.5, largely structural: on the present stimuli, sharing and separation imply almost the same predictions, so human responses cannot adjudicate between them. Separating the two would require a different stimuli design featuring generalisation probes placed between an exception and the family it resembles, where pooled and separate shapes imply different forms, or exceptions phonologically far enough from

that family that pooling them with the prototype would meaningfully shift the cluster's predictions. We therefore use the shared model only to read off the structure it induces, and leave to future work where cross-class co-motivation matters, and for which systems.

5.2 Representations are dynamic

The three models differ most on their predicted trajectory for the exceptions. Successful learners' performance on exceptions changes over the course of the experiment, at first they are treated as if they belonged to the same class as their phonological prototype, then the suffix choice switches to match the intended one. As the item does not change over the experiment, the cause of the difference in responses must be found within the learner.

From a mentalist perspective, linguistic behaviour is the result of applying a deployment process to mental representations - a change in behaviour must therefore be the result of a change in deployment process or a change in representations. The three models we present all have the same deployment mechanism (i.e. the prediction process described in Section 3.2.3), and differ in their representations, so within our setup this is where a change must occur in order for behaviour to change. Representations, in turn, can shift in two ways: they can refine their content, sharpening their estimates while maintaining a fixed structure, or they can change their structure, lumping and splitting categories. The exemplar and prototype model can only do the former. It is the second kind of change, the reshaping of structure, that appears to be necessary to model the trajectory of the exceptions. The prototype model reduces its suffix class to the average shape of its members, and because the exception resembles the members of the other class, it ends up with the wrong suffix. The exemplar model correctly associates the exception's trace with its suffix, but has a hard time acting on this information because the trace contributes only through similarity, and the exception's own traces are outvoted by members of its phonological prototype, which take the wrong suffix. Accumulating traces for the exception as the experiment progresses do not provide enough evidence, because the phonological neighbours gain traces at the same rate. To get the exception to behave as intended, one needs to set it apart from the words it phonologically resembles, recarving the space, which is what the adaptive model uniquely allows for in the set considered: the exception forms its own cluster once enough evidence builds up.

This framing provides a different reading key for a familiar phenomenon. Children overextend a dominant inflection before they master its exceptions (Bybee & Slobin, 1982; Marcus et al., 1992). This is usually described in terms of a new rule entering their grammar - within the current account, the phenomenon is the result of the lexicon being restructured to a finer grain, to separate the exceptions from the bigger cluster they had been lumped into, being able to serve as their own separate reference class.

Locating any change in behaviour within a change in representation in our models is a design decision: we let the representation vary and held fixed the deployment process, so any change the models can express is necessarily the result of a change in representations. Proponents of, for example, exemplar models might reasonably object

that exemplar representations are compatible with a change in behaviour on exceptions by having a deployment mechanism that learns to discount unhelpful neighbours. This would be a fair critique: representations and deployment process trade off against each other, and behaviour cannot serve as evidence for one without holding the other still (Anderson, 1978). This would salvage the exemplar model, but it would do so through the same high-level mechanisms posited for the adaptive model, ultimately changing the reference set for a particular lexeme, by narrowing in on the correct set as evidence grows. The behavioural evidence discussed here and elsewhere shows uncontestably that this narrowing takes place in step with the amount of evidence. The change in behaviour itself does not settle whether this is a change of representations or of deployment process. In the modeling presented in this paper, we focused on changes in the representations, partly in reaction to a tendency in theories of the lexicon to load the whole explanatory burden onto the representation and leave the deployment process underexamined. This choice should remain something we have to argue for, as the behaviour does not determine it. The same is true for any account that gets its evidence from corpora, experiments, or other records of what speakers do.

5.3 What to do about this

The picture of representations sketched here, not a fixed inventory but something inferred and updated, may look radical to morphological theory. This is however the long-accepted norm in the cognitive science of concepts, which gave up stable, retrieved category structure decades ago: graded structure is unstable across contexts and speakers, and concepts behave less like stored objects than like temporary constructs assembled to fit the moment (Barsalou, 1987; Yee & Thompson-Schill, 2016). This change did not leave the field of cognitive science adrift, instead it proved generative. Once a concept is taken to be assembled for the occasion rather than retrieved intact, the question shifts from what is stored to how a representation is built to fit the situation, and cognitive science has pursued that question along several lines: grounded and situated cognition, where a concept is a context-tailored perceptual simulation (Barsalou, 2008); ad hoc categories, constructed on the fly to serve a current goal (Barsalou, 1983); controlled semantic cognition, where a control network shapes what is retrieved to suit the context (Lambon Ralph, Jefferies, Patterson, & Rogers, 2017; Yee & Thompson-Schill, 2016); and the rational and Bayesian modelling of category learning, where the structure of a category is inferred from the evidence rather than stipulated (Anderson, 1991; Griffiths et al., 2007). It is this last strand we draw on, in its resource-bounded form (Lieder & Griffiths, 2020), which treats a representation as the best a learner can do with the evidence at hand and reasons from the problem toward the mechanism. Positing dynamic categories is not entirely novel in its application to morphological data: Ambridge's (2020a; 2020b) exemplar account builds generalisation on the fly rather than storing it, and the discriminative-learning tradition derives morphological behaviour from experience without a fixed inventory of stored abstractions or paradigms (Baayen & Ramscar, 2015). The ingredients for a dynamic, learning-driven morphology are already in the field. None of this compels morphology to become a theory of learning. But a theory that claims cognitive reality, or that tests itself against human behaviour, has already taken on the commitment

these traditions make good on: that a representation and the process that deploys it are, together, the predictors of what learners do.

Taken to its logical conclusion, this approach also yields expectations we have not tested here, and offer only tentatively. Because grain tracks evidence, the high-frequency regions of a lexicon might be expected to individuate before its low-frequency tail, fine distinctions surfacing first where exposure is densest, and because the evidence has to support the distinctions drawn, a system in which linguistic structure (phonology, semantics, or paradigmatic information) reliably cues class membership might sustain finer grain, at a given amount of exposure, than one in which it does not.

6 Conclusion

Morphology has treated the level of abstraction a speaker represents as a fact about architecture. We argue instead that this is a quantity the learner infers, set by the evidence available, and revised as that evidence grows. Our participants' representations are best captured not by a single grain but by one that shifts from coarse to fine as they gather more evidence. We suggest that the question "which level of abstraction?" should be traded for one with a moving, measurable answer: what does the evidence, at this point in learning, warrant?

References

Ackerman, F., Blevins, J.P., Malouf, R. (2009, 07). 3 parts and wholes: Implicative patterns in inflectional paradigms. *Analogy in grammar: Form and acquisition.* Oxford University Press. Retrieved from https://doi.org/10.1093/acprof:oso/9780199547548.003.0003

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, *89*(3), 429–464, https://doi.org/10.1353/lan.2013.0054

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161, https://doi.org/10.1016/S0010-0277(03)00146-X

Ambridge, B. (2020a). Abstractions made of exemplars or 'you're all right, and i've changed my mind': Response to commentators. *First Language*, *40*(5–6), 640–659, https://doi.org/10.1177/0142723720949723

Ambridge, B. (2020b). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5–6), 509–559, https://doi.org/10.1177/0142723719869731

Anderson, J.R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*(4), 249–277, https://doi.org/10.1037/0033-295x.85.4.249

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. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233,

Ashby, F.G., & Maddox, W.T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178, https://doi.org/10.1146/annurev.psych.56.091103.070217

Baayen, R.H., Dijkstra, T., Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, *37*, 94–117,

Baayen, R.H., & Ramscar, M. (2015). Abstraction, storage and naive discriminative learning. E. Dąbrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 100–120). Berlin: De Gruyter Mouton.

Baer-Henney, D., Kügler, F., van de Vijver, R. (2015). The interaction of language-specific and universal factors during the acquisition of morphophonemic alternations with exceptions. *Cognitive Science*, *39*(7), 1537–1569, https://doi.org/10.1111/cogs.12209

Barsalou, L.W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227, https://doi.org/10.3758/BF03196968

Barsalou, L.W. (1987). The instability of graded structure: Implications for the nature of concepts. U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge University Press.

Barsalou, L.W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645, https://doi.org/10.1146/annurev.psych.59.103006.093639

Beniamine, S., Bonami, O., Luís, A.R. (2021). The fine implicative structure of European Portuguese conjugation. *Isogloss*, *7*(9), 1–35, https://doi.org/10.5565/rev/isogloss.109

Blevins, J.P. (2008). Declension classes in Estonian. *Linguistica Uralica*, *44*(4), 241–267,

Blevins, J.P. (2016). *Word and paradigm morphology*. Oxford: Oxford University Press.

Blevins, J.P., Milin, P., Ramscar, M. (2017). The Zipfian paradigm cell filling problem. F. Kiefer, J.P. Blevins, & H. Bartos (Eds.), *Perspectives on morphological organization: Data and analyses* (pp. 139–158). Leiden: Brill.

Bonami, O., & Beniamine, S. (2016). Joint predictiveness in inflectional paradigms. *Word Structure*, *9*(2), 156–182, https://doi.org/10.3366/word.2016.0092

Booij, G. (2010). *Construction morphology*. Oxford: Oxford University Press. Retrieved from https://global.oup.com/academic/product/construction-morphology-9780199571918

Brooks, P.J., Braine, M.D.S., Catalano, L., Brody, R.E., Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language*, *32*(1), 76–95, https://doi.org/10.1006/jmla.1993.1005

Brown, D., & Hippisley, A. (2012). *Network morphology: A defaults-based theory of word structure*. Cambridge University Press.

Butterworth, B. (1983). Lexical representation. B. Butterworth (Ed.), *Language production: Vol. ii. development, writing and other language processes* (pp. 257–294). London: Academic Press.

Bybee, J. (2001). *Phonology and language use*. Cambridge University Press.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press. Retrieved from https://books.google.co.uk/books?id=Htwy_iPG114C

Bybee, J., & Moder, C.L. (1983). Morphological classes as natural categories. *Language*, *59*(2), 251–270, https://doi.org/10.2307/413574

Bybee, J., & Slobin, D.I. (1982). Rules and schemas in the development and use of the English past tense. *Language*, *58*(2), 265–289, https://doi.org/10.2307/414099

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54, https://doi.org/10.18637/jss.v100.i05

Caramazza, A., Laudanna, A., Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, *28*, 297–332,

Corbett, G.G. (2009). Canonical inflectional classes. F. Montermini, G. Boyé, & J. Tseng (Eds.), *Selected proceedings of the 6th décembrettes* (pp. 1–11). Somerville, MA: Cascadilla Proceedings Project.

Culbertson, J., Gagliardi, A., Smith, K. (2017). Competition between phonological and semantic cues in noun class learning. *Journal of Memory and Language*, *92*, 343–358, https://doi.org/10.1016/j.jml.2016.08.001

Dekker, P., Rasilo, H., de Boer, B. (2025). *Learning inflection classes using adaptive resonance theory.* Retrieved from https://arxiv.org/abs/2512.15551

Divjak, D., & Milin, P. (2026). The synergy of linguistic theories and models of learning. *Linguistics Vanguard*, , https://doi.org/10.1515/lingvan-2024-0243 (Advance online publication)

Dreyer, M., & Eisner, J. (2011, July). Discovering morphological paradigms from plain text using a Dirichlet process mixture model. *Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP)* (pp. 616–627). Edinburgh, Scotland, UK. Retrieved from https://aclanthology.org

Frigo, L., & McDonald, J.L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*(2), 218-245, https://doi.org/https://doi.org/10.1006/jmla.1998.2569 Retrieved from https://www.sciencedirect.com/science/article/pii/S0749596X98925694

Goldberg, A.E. (2006). *Constructions at work: The nature of generalization in language.* Oxford, UK: Oxford University Press.

Goldberg, A.E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions.* Princeton: Princeton University Press. Retrieved 2026-06-23, from https://doi.org/10.1515/9780691183954

Goldwater, S., Griffiths, T.L., Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54, https://doi.org/10.1016/j.cognition.2009.03.008

Griffiths, T.L., Canini, K.R., Sanborn, A.N., Navarro, D.J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. *Proceedings of the annual meeting of the cognitive science society* (Vol. 29). Retrieved from https://escholarship.org/uc/item/5p6505mf

Hay, J.B., & Baayen, R.H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences*, *9*(7), 342–348, https://doi.org/10.1016/j.tics.2005.04.002

Hilpert, M. (2021). *Ten lectures on diachronic construction grammar*. Leiden, The Netherlands: Brill. Retrieved from https://brill.com/view/title/56854

Jackendoff, R., & Audring, J. (2019). *The texture of the lexicon: Relational morphology and the parallel architecture*. Oxford: Oxford University Press.

Kam, C.L.H., & Newport, E.L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195, https://doi.org/10.1080/15475441.2005.9684215 Retrieved from https://doi.org/10.1080/15475441.2005.9684215

Kapatsinski, V. (2018). *Changing minds changing tools: From learning theory to language acquisition to language change*. Cambridge, MA: MIT Press.

Lambon Ralph, M.A., Jefferies, E., Patterson, K., Rogers, T.T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55, https://doi.org/10.1038/nrn.2016.150

Lieder, F., & Griffiths, T.L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1, https://doi.org/10.1017/S0140525X1900061X

Marcus, G.F., Pinker, S., Ullman, M., Hollander, M., Rosen, T.J., Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), , (Serial No. 228)

Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238, https://doi.org/10.1037/0033-295X.85

.3.207

Nosofsky, R.M. (1986, March). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–61, https://doi.org/10.1037//0096-3445.115.1.39

O'Donnell, T.J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. Cambridge, MA: MIT Press.

Palancar, E.L. (2014). Revisiting the complexity of the chinantecan verb conjugation classes. J.-L. Léonard & A. Kihm (Eds.), *Patterns in mesoamerican morphology* (pp. 77–102). Paris: Michel Houdiard.

Perfors, A., Tenenbaum, J.B., Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(3), 607–642, https://doi.org/10.1017/S0305000910000012

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193, https://doi.org/10.1016/0010-0277(88)90032-7

Reed, S.K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407, https://doi.org/10.1016/0010-0285(72)90014-X

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233, https://doi.org/10.1037/0096-3445.104.3.192

Round, E.R., Esher, L., Beniamine, S. (2024). The natural stability of autonomous morphology: How an attraction–repulsion dynamic emerges from paradigm cell filling. *Morphology*, , https://doi.org/10.1007/s11525-024-09433-y

Rumelhart, D.E., McClelland, J.L., PDP Research Group. (1986). *Parallel distributed processing* (Vol. 1). The MIT Press.

Schreuder, R., & Baayen, R.H. (1995). Modeling morphological processing. L.B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 131–154). Erlbaum.

Schuler, K.D., Yang, C., Newport, E.L. (2016). Testing the tolerance principle: Children form productive rules when it is more computationally efficient to do so. *Proceedings of the annual meeting of the cognitive science society* (Vol. 38). Retrieved from https://escholarship.org/uc/item/1cn5n52p

Stump, G., & Finkel, R.A. (2013). *Morphological typology: From word to paradigm* (No. 138). Cambridge: Cambridge University Press.

Taft, M., & Forster, K.I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning & Verbal Behavior*, *14*(6), 638–647, https://doi.org/10.1016/S0022-5371(75)80051-X

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581,

Ullman, M.T. (2001). The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research*, *30*(1), 37–69, https://doi.org/10.1023/A:1005204207369

Ullman, M.T. (2004, May-June). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*(1-2), 231–270, https://doi.org/10.1016/j.cognition.2003.10.008

Wonnacott, E., Newport, E.L., Tanenhaus, M.K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*(3), 165–209, https://doi.org/10.1016/j.cogpsych.2007.04.002

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. Cambridge, MA: MIT Press.

Yee, E., & Thompson-Schill, S.L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, *23*(4), 1015–1027, https://doi.org/10.3758/s13423-015-0948-7 Retrieved from https://link.springer.com/article/10.3758/s13423-015-0948-7

Declarations

- Competing Interests

  The authors have no competing interests to declare.

Supplementary Files

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

- zenododeposit.zip