An Incremental Program Induction Model of Slow Mapping Words to Meanings

Ella Markham (e.markham@sms.ed.ac.uk)
School of Informatics, University of Edinburgh

Hugh Rabagliati
Department of Psychology, University of Edinburgh

Neil R. Bramley
Department of Psychology, University of Edinburgh

Abstract

The process by which people adjust, enrich, and revise their understanding of word meanings over time – so-called 'slow mapping' – has often been overlooked, particularly in terms of how a computationally bounded learner might approach it. To address this gap, we propose a process model of incremental word-meaning induction. This proposal is inspired by recent work on concept and theory change grounded in a probabilistic language of thought (pLOT). We focus on the problem of fixing the meanings of words from usage examples, taking kinship terms as our initial test domain. We frame word meaning induction at a computational level as a program induction problem, and hypothesize that individual learners search for possible meanings as evidence arrives via a mutative Markov-Chain Monte-Carlo search scheme. We show this idea provides a better description of how participants' generalizations and definitions of alien kinship words shift as evidence arrives, outperforming normative accounts and other baselines.

Keywords: slow mapping; word learning; incremental model; pLOT; tree surgery

Introduction

In the noisy language learning environment, we must gradually refine our understanding of word meanings as we hear them applied to many referents and across many contexts. This gradual process has been termed 'slow mapping', in contraposition to the phenomenon of 'fast mapping' where learners make an initial guess of a word meaning based on little evidence (Carey & Bartlett, 1978). The computational processes through which word meanings are shaped and adjusted remain surprisingly mysterious, despite their centrality to language learning.

One set of tasks that have examined slow mapping is Cross-Situational Word Learning (CSWL). In these tasks, learners must triangulate the meaning of a word via its co-occurrence with different objects across trials. Various theories have been proposed as to the strategies of learners in these settings. Some authors (e.g., Yu & Smith, 2007), have modeled learners as solving the problem ideally: storing all co-occurrence statistics, enumerating all possible word-meaning hypotheses and selecting the maximum a posteriori (MAP) hypothesis for each word. Meanwhile, Propose-but-Verify (PbV) accounts (Medina et al., 2011; Trueswell et al., 2013) treat learners as doing something much simpler: storing only a single hypothesis about a word meaning at a time. Upon encountering evidence it cannot account for, this is discarded and a new hypothesis, consistent with the focal evidence, is resampled. On the face of it, both these proposals fall short of plausibility for describing word learning beyond the lab. In general, one cannot enumerate the infinite set of meanings a word could have, nor store words' co-occurrence statistics globally and indefinitely. On the other hand, discarding hard-won word meanings entirely when they are challenged and regenerating based on current evidence also likely will not work in general. Hence, learners need a way to generate revised hypotheses while having available only a compressed 'semanticised' representation (Tulving, 1972). A third path seems necessary in which meanings are refined by evidence rather than discarded, but where these refinements take place without recourse to a global posterior distribution or complete episodic memory.

While PbV may not scale beyond CSWL tasks, evidence supporting such hypothesis-testing accounts has come from the prevalence of order effects (Medina et al., 2011; Trueswell et al., 2013). Normatively, the order one experiences evidence should not affect one's final beliefs. However, semantic learning tasks often produce order effects, such as in colour category (Sandhofer & Doumas, 2008), causal structure (Bramley et al., 2017), and concept learning (Zhao et al., 2024). 'Garden pathing' order effects are also prevalent in fast online meaning comprehension involved in sentence processing (Frazier & Rayner, 1982). Markham et al. (2024) found that word learning in the kinship domain is also subject to order effects, making this a suitable domain to try to understand the process of word-meaning revision. That is, a satisfying account of slow mapping should explain the influence of order of learning experiences on the meanings learners arrive at.

A Mutative Model of Slow Mapping

We present a computational model of slow mapping that, like PbV, assumes that learners store a single meaning hypothesis for each word. However, unlike PbV, when confronted with new data that is incompatible with a word meaning hypothesis, it generates mutations of that hypothesis, tending to accept those that improve the fit (i.e., explain the new observation), while also improving or retaining the degree of parsimony (the complexity of the definition).

Because the mutations proposed are syntactically 'local', this proposal diverges from the computational-level Bayesian ideal of word learning in which meaning change purely is based on posterior probability (fit and parsimony). Given that

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

the revised hypotheses a learner considers depend on their previous guesses, our account will, in general, produce order effects if its mutative search is short (as well as when older data is forgotten), asymptoting to posterior sampling if its search is sufficiently long (provided older data is remembered).

This kind of incremental revision has deep roots in the cognitive science literature. In philosophy of science, the Duhem-Quine thesis is that conceptual change is always limited by a theorist's preexisting assumptions (Quine, 1969), while Lakatos (1970) highlights that scientists often accumulate auxiliary hypotheses and exceptions to maintain their theories in the light of contradictory evidence, rather than discarding them completely. Of particular relevance is a family of approximate Bayes-inspired models which capture anchoring effects as the result of bounded approximation to rational inference via local sampling processes (Lieder et al., 2018; Dasgupta et al., 2017; Sanborn & Chater, 2016; Bramley et al., 2017, 2023; Fränken et al., 2022).

A Hypothesis Space as a pLOT

We conceptualize a latent mental hypothesis space of potential word meanings as a 'probabilistic language of thought' (pLOT) relating words to their extensions in logical compositions of features. A LOT (Fodor, 1975) is a set of conceptual primitives and the rules to combine them, expressing an open set of concepts that can be expressed by combination of primitives. A PCFG (Piantadosi & Jacobs, 2016) defines a set of iterative productions that make it possible to randomly sample all legal expressions in the LOT. Given that each production occurs with a specific probability, the chance of drawing a particular hypothesis from a PCFG can be equated with its 'prior' probability, with more complex hypotheses (composed of more primitives) tending to have lower prior probabilities than simpler ones. pLOTs have been widely used as a computational abstraction for defining normative inference in open-ended hypothesis spaces such as concepts (Goodman et al., 2008; Piantadosi et al., 2016) and in this case, word meanings. Generically we can define the posterior over word meaning hypotheses $h \in H$ as a Bayesian combination of its PCFG-defined prior and the likelihood of each hypothesis producing our observations of that word usage $\mathbf{d}$:

$$P(H|\mathbf{d}) \propto P(\mathbf{d}|H)P(H). \quad (1)$$

In the word meaning domain, the likelihood of a word being used to refer to an object can be complex and context-specific, but as a reasonable starting point the *size principle* (Xu & Tenenbaum, 2007) captures that general words spread their likelihood across many different referents while specific words concentrate it on the few things they apply to.

Unfortunately, it is very hard to approximate Equation 1 because set H is generally infinite. A naive approach is to generate samples from the PCFG prior and weight these by their likelihood. However, nontrivial hypotheses are extremely rare to generate a priori. A more workable approach, typically used in practice, is to do Markov Chain Monte Carlo (MCMC) to explore the posterior directly, generating a series of autocorrelated posterior samples, each generated by proposing local mutations to the previous one, with the help of the PCFG but accepted as a function of their unnormalised posterior probability (i.e., likelihood$\times$prior) relative to the previous sample. This allows more of the samples considered to inhabit high probability regions of the posterior even if these are too low prior to be feasibly generated a priori.

With a long enough sampling chain the endpoint is a posterior sample, and the most frequently visited hypothesis is likely to be the true MAP. While extensive sampling is necessary to generate normative predictions, this is normally couched at the computational level (Marr, 1982). However, some recent work leans into the idea that such a mechanism also doubles as a process level hypothesis about how cognisers might arrive, if not at the MAP solution, at least at a revised hypotheses that is shaped by considerations of both parsimony and fit. While a long re-sampling chain can approach independent posterior sampling, it can often be boundedly rational to accept some auto-correlation in return for performing less inferential work (Lieder et al., 2018; Vul et al., 2014). Moreover, to the extent that one's starting hypothesis summarizes—or in this cases 'semanticizes'—data one has now forgotten, anchoring on one's current hypothesis can actually play a prior-like role (Bramley et al., 2017). Following this logic, several papers have modeled order-effects as the result of short MCMC-like chains of proposed-and-accepted mutations, initialized at the learner's previous hypothesis. We follow such a framework to test our own theory.

In order to compare our model predictions to behavioral data, we ran a new experiment where participants learn kinship terms across multiple exposures, replicating and improving on the experiment in Markham et al. (2024). The model and experimental set-ups are described in detail below.

Modeling Framework

Grammar Whilst there are many possible grammars to express kinship, we selected a set of primitives that allow us to refer to features including gender, generation and ancestral relationships, alongside the booleans with which to combine them.

In order to express chains of relations, we included a primitive 'chain' function that returns true if the relevant intermediate family members exist. For example, 'the sister of my mother' (i.e., maternal aunt) can be expressed as *chain([sister,mother],y,X)*. A limitation of our assumed grammar is that gender and parenthood cannot be expressed simultaneously within the chain function. Therefore we include relational primitives which are the combination of such features (e.g., sister).

Given that our proposed model implements local changes and a key form of this is a localised elimination, we included null primitives that functionally eliminate what they replace (e.g., altering 'older female relative' to just 'older relative').

Concretely, we include *true(y,X)* and *false(y,X)* which, when entered in a conjunction or disjunction, respectively result in the truth of a boolean relying wholly on its other element. We also include *identity(y,X)*, true iff y and X are equal. Replacing a relational term in a chain with this functionally eliminates that relational step while preserving the rest of the chain.

The full grammar is shown in Figure 1a.

Prior For simplicity, we approximate the prior probabilities for word meanings by the number of nodes within the tree structure of that hypothesis:

$$P(h) \approx \left[\frac{1}{|h|}\right]^{1/t} \tag{2}$$

Where t is a temperature parameter we later fit with a rough grid search.

Likelihood For each hypothesis, we assumed a non-deterministic likelihood function, calculated as the sum of the likelihoods for each example shown, accounting for the size principle (Xu & Tenenbaum, 2007). Fitted parameter ε captures the possibility of 'misspeaking' errors, where a word is used to refer to someone that it does not actually refer to.

$$p(y|h) = \sum_{e \in E} \frac{\varepsilon}{m} + \frac{\sigma(1-\varepsilon)}{\text{size}(h)} \tag{3}$$

Where σ represents whether or not the example is part of the hypothesis extension and can take values $[0,1]$, and m represents the number of members on the tree minus the speaker.

Tree Surgery as Local Adaptation

In the current work, we propose a similar MCMC-based local search model to the Fränken et al. (2022) model of rule-based concept induction, adapted to the problem of inferring word meanings. In our task, participants reason about the definitions of alien kinship terms as they see them in use. We selected kinship as our domain of study given it has been a domain of recent focus and admits an intuitively compositional structure making it a natural initial test bed for these ideas. Research in this area has previously used a pLOT to examine the informational properties of different kinship systems (Kemp & Regier, 2012; Mollica & Piantadosi, 2019).

Following Fränken et al. (2022), we propose that when probed, participants produce a best guess hypothesis about the meaning of a kinship word based on the examples they have seen via MCMC-like local search. Where they have a previous best guess and must integrate a new usage, they start a short MCMC chain from their previous guess to update the meaning.

We assume a proposal distribution inspired by 'Tree Regrowth' (TR) (Goodman et al., 2008) and 'Tree Surgery' (TS) (Fränken et al., 2022). Tree Regrowth involves selecting a node at random from a hypothesis, deleting everything underneath it, and regrowing a new branch from that node using the PCFG. This allows for locally similar hypotheses to be generated in multiple ways, including the deletion of a branch or a relation in a chain, with the replacement of a node with *true/false* and *identity* respectively. However, while this provides some flexibility in the possible local moves, participants may also choose to keep and build upon their current hypothesis (i.e., by adding on an auxiliary hypothesis). Therefore, we include Tree Surgery, which involves the addition of a conjunction or disjunction to the current hypothesis, with the other element of the conjunction or disjunction being regrown from a node selected at random from the current hypothesis (see Figure 1 below). This results locality in multiple respects. For example, taking the hypothesis 'older female relative', expressible in our grammar as $\wedge(\text{female}(y), \text{generation}+2)$, a tree-regrowth move might replace female with A, generating the new branch replacing A until termination, either replacing or eliminating it. Alternatively, a conjunction or disjunction might be added, e.g. 'older female relative or sister'. At each step in the chain, there is an equal probability of adapting the hypothesis via TR or TS. If TS is selected, we assume an equal probability of adding on a conjunction or a disjunction. Example moves are illustrated in the Figure 1.

To decide whether to accept the new regenerated hypothesis as the next step in the chain, we used a Metropolis-Hastings algorithm, with the hypothesis accepted with probability

$$\min(1, \frac{P(\mathbf{d}|h')P(h')}{P(\mathbf{d}|h^t)P(h^t)}) \tag{4}$$

Where $P(h^t)$ is the prior probability of the current hypothesis, $P(\mathbf{d}|h^t)$ is the likelihood of the examples under t and similarly for proposed hypothesis h'.[1]

Experiment

Participants have to learn the meanings of a series of alien kinship terms by exposure to their usages, and must repeatedly make generalizations about which other kin can be referred to by the word. Participants learn three words spoken by three different aliens (each of whom speaks a different alien language). For each trial, the aliens use a word to refer to a specific family member, who is highlighted on the family tree. Participants are then asked to select everyone on the family tree that they believe can be referred to using that word.

In addition to the Markham et al. (2024) task, between each word exposure, participants engage in a distractor task. They are shown a new 'alien' object which they will have to recall

[1]The Metropolis-Hastings algorithm normally includes a correction for detailed balance: $p(h^t|h') = p(h'|h^t)$. If not met, as in our case, this should be factored into the acceptance function. However, the correction is infeasible to calculate in this case, given the very large number of possible regeneration pathways between many hypothesis-pairs. Therefore, we do not include this in our acceptance function. While we leave the broad consequences for future study (cf. Suwa & Todo, 2010), we note that an MCMC-like mechanism without detailed balance may act similarly to a human learner in that both struggle to escape local minima.

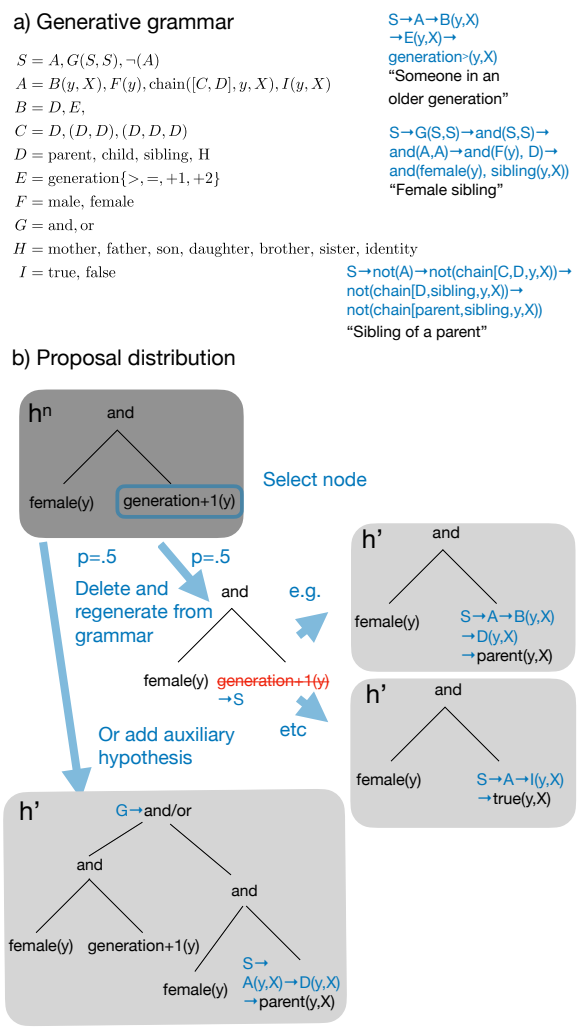

Figure 1: (a) PCFG grammar. (b) Proposal distribution: h^n can be mutated by random selection of a node and either deleting and regrowing it from the PCFG, or by splicing in a conjunctive or disjunctive node.

later. We include this task to simulate that, in reality, word exposures tend to be spread out and interpolated with other tasks, limiting the extent to which participants can maintain evidence in working memory.

Following Medina et al. (2011), we assume that participants store a hypothesis of the word meaning. Therefore, we test sequences of examples that we expected to 'garden path' participants towards a certain hypothesis, before introducing a final example that does not fit with this hypothesis. This allows us to create moment of dramatic change in the posterior in which we anticipate that the hallmarks of incremental local adjustment will be easy to spot. We also contrast with conditions in which participants see the same data in the reverse order, yielding the same normative predictions but different localist predictions, again helping us tease apart the rational and process level considerations.

Methodology

Participants 200 UK or US based adults were recruited via Prolific (92 female; 1 no response, Age (median): 36, Range: 19-74, Prolific approval rate $\geq 99\%$). We excluded and replaced participants whose written answers indicated non-engagement (i.e., were nonsensical or unrelated to the kinship family tree) (N=9) and who never included any family members that had been referred to in their selections (N=3).

Figure 2: Family tree shown to participants. Letter labels $a-t$ were not shown to participants. The person being referred to by the alien is highlighted with a red circle.

Procedure Participants were told that they had arrived on an alien planet and had to decipher the words that the aliens used to describe their family members, based on a series of examples. Participants learned a practice word ('immediate family') prior to starting the task to ensure understanding of the task interface.

For each trial, the alien would refer to a member of their family four times. Each time, the person they were referring to was highlighted on the tree. After each example, participants were required to select everyone on the family tree that they believed could be referred to using that word (and deselect everyone they believed could not be referred to using that word). They then gave a rating $\in (1,5)$ reflecting their confidence. At the end of the trial, they gave a written guess of what they believed that the word meant. They were then given the opportunity to change their selection based on this written guess.

Simultaneous with the main task, in order to mimic the separation of real word usages, there was a distractor task. After every word use, participants were shown an object associated with that alien that they had to remember until the end of the trial. Following the final example for each word, participants were asked to select the items they had seen associated with that alien out of a set of ten alien objects.

During the instructions, it was made clear to participants that the languages spoken by the aliens were distinct, both from each other and also from English. Participants were also informed that they would receive a bonus of up to £0.50 dependent on their performance in both the family tree selection and memory tasks. In reality, as long as they completed both tasks properly, they received a bonus. They also performed

Table 1: Model Predicted Guesses by Word & Condition

Word 1	MAP	MAP (reverse)	Local
1 (4) k	"female"	"female"	("female")
2 (3) i	"aunt"	"female"	
3 (2) k	"aunt"	"female"	
4 (1) g	"female"	"female"	("aunt or grandmother")

NOTE: Letters correspond to the family members as shown in Figure 2, parentheses show the reversed order.

a comprehension check where they were asked to select specific family members (e.g., 'mother') to ensure their ability to understand a family tree.

Design and Stimuli We tested three words, each of which involved four example usages, and we varied the order of these usages (forward/reverse) between subject. The stimuli in the forward condition were designed to induce a dramatic shift in the MAP hypothesis between the penultimate and final trial. If participants were not selecting the MAP hypothesis and instead following an anchored local model, such as ours, we would expect this to provide a good opportunity to observe order effects, with differences between local and global model predictions and differences in the local model predictions between the forward and reverse conditions.

To induce expected differences, participants in the forward condition were repeatedly shown similar members (e.g., aunts) on the tree to 'garden path' them towards a a specific hypothesis having the highest posterior probability on the penultimate trial $T-1$ (e.g., 'aunt'). They were then shown a family member who does not fit with this hypothesis at trial T (e.g., paternal grandmother). According to our incremental model, they should then make a local change to their hypothesis at $T-1$ to accomodate this data (e.g., 'aunt or grandmother'), resulting in a hypothesis that is different to the MAP (e.g., 'female'). Such a local move should not be possible from $T-1$ in the reverse condition, leading to expected differences between the two conditions. A demonstration of this is shown in Table 1.

The family tree contained 20 family members of 3 generations, alongside the alien speaker (see Figure 2). Female and male members were pink and blue respectively. Ellipses were used to indicate members that were part of the tree only by marriage and would not be referents of the words used. The items used for the memory distraction task were taken from Horst & Hout (2015).

Results

Prior to model fitting, we tested for for order effects by examining final selections (Figure 3). For each word type, we create categories for selection patterns made by ≥ 5 participants (across both conditions), while classifying any other selections as 'other'. For all 4 words, there was an effect of order on the final selections made by participants: Word 1 ([k,i,k,g]) $X^2(8, N = 200) = 38.6, p < .001$), Word 2 ([s,t,s,c]) $X^2(6, N = 200) = 25.7, p < .001$) and Word 3 ([f,f,f,j]) $X^2(8, N = 200) = 17.8, p = .02$).

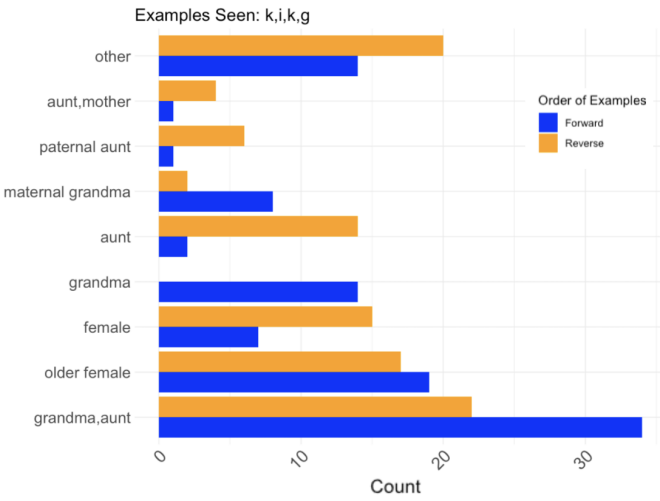

Figure 3: Participant selections for the Word 1 ([k,i,k,g]) final trial for both forward and reverse conditions. See osf.io/nd4j2/ for full data and figures for Words 2 and 3.

These order effects support an account of slow mapping which involves making a local change. We examine this at an individual level with the help of our incremental meaning model.

Model Fitting

We fit four models to the data, allowing us to compare the predictions of local vs alternative accounts of slow mapping.

Random Baseline We assume participants' selections are completely random (each family member is selected with probability .5).

Normative Baseline We generate a sample of 50,000 hypotheses from our pLOT, over which we calculate the normalised posterior distribution using the prior and likelihood described above.

In order to accommodate response noise, we introduce a likelihood function mapping from hypotheses to selections $\mathbf{s}$, which allows for some misclicks:

$$P(s_y|h) \propto \exp(-N_{\text{misclicks}}) \tag{5}$$

Hence, $P(\mathbf{s}|\mathbf{d})$ is given by:

$$P(\mathbf{s}|\mathbf{d}) = \sum_{h \in H} P(\mathbf{s}|h)P(h|\mathbf{d})P(h) \tag{6}$$

No Change Baseline To account for the fact that participants may maintain a singular hypothesis across trials, we include a baseline that predicts participants will maintain their

Table 2: Model Fitting Results

Model	t	ε	BIC	N Best Fit
Random	-	-	72,000	0
Normative	0.6	0.1	9,508	12
No Change	-	-	8,770	33
Local Model	0.6	0.1	**5,593**	**155**

selection from the previous trial with any deviations just random noise.

$$P(s|s^{t-1}) = \exp(-N_{\text{misclicks}}) \tag{7}$$

Local Model The local model predicts that new hypotheses are generated via MCMC-like TS methods and anchored to their previous hypothesis (i.e., the chain is started on the previous hypothesis). Therefore, the new and old hypotheses should show syntactic similarities. Hence, to simulate this dependency, we ran 10,000 very short chains starting from the simplest syntactic representation of the participants' current hypothesis as derived from their previous selection. Each chain was terminated once a new hypothesis was accepted.

Incorporating the error term used for the normative baseline, the marginal probability of each potential new hypothesis given the data and previous hypothesis was calculated as:

$$P_{TS}(h|h^{t-1}, \mathbf{d}) = \frac{\sum(h)}{N} \tag{8}$$

With $P(\mathbf{s}|\mathbf{d})$ given by:

$$P(\mathbf{s}|\mathbf{d}; h^{t-1}) = \sum_{h \in H} P(\mathbf{s}|h) P_{TS}(h|h^{t-1}, \mathbf{d}) \tag{9}$$

Model Fitting Results

The fit for models on all trials is shown in Table 2. The local model has a much better fit to the data compared to all other models, while also being the most accurate model for the highest number of participants.

Discussion

In this paper, we introduced a new model of slow mapping words to meanings, in which learners accommodate new evidence by making local and incremental changes (via Tree Surgery) to their current hypothesis of the word meaning. We fit this model, alongside baselines, to data gathered in a kinship term learning task, by running short MCMC-like chains anchored on the participants' previous hypotheses. We found that our incremental model outperformed our baseline models. Hence, the model provides a new and promising perspective from which to examine the computational process of slow mapping, accounting for the individual pathways that learners may take as they navigate the word learning environment.

These results provide a strong foundation to build further model adjustments upon. For example, examination of the model predictions shows that a strength of the local model on less 'surprising' trials is its strong bias towards maintaining the current hypothesis, a behaviour also shown in participants. The 'No Change' baseline also has this behaviour, however the local model is much better at accounting for the adjustments made on 'surprising' trials, striking a better balance between maintaining and altering hypotheses. Nevertheless, while the local model is able to better move between maintaining and altering hypotheses than our other baselines, the local model is often overly biased (in comparison to participant data) to maintain a hypothesis on surprising trials. This is most likely because the local model stores and calculates over all the examples shown, the majority of which, in our paradigm, fit with the current hypothesis. Therefore, a local model with increased memory constraints may fit even better to the data. Overall, such results also indicate that perhaps Win-Stay Lose-Sample models (WSLS, Bonawitz et al., 2014), where the sampling involves local search (potentially with limited memory), may also be viable process-level candidates. See osf.io/nd4j2/ for further details and data specifics.

Although the current work focuses on kinship terms, our model can be extended to account for learning across other word domains. Given the biases that native English speakers will bring to a kinship learning task, further experiments in 'nonsense' domains in which participants bring very limited language prior would provide an interesting extension to the work presented here. Although they may not be able to be represented using a pLOT, words that follow more of a 'prototypicality' structure seem likely to also be mapped through a process of anchoring and incremental change and hence, should be examined further in future work.

Further research should also consider how learners select the primitives that their grammar or lexicon consists of to begin with. Our current model already assumes a set of primitives that the learner comes equipped with, as do others in the literature (Fränken et al., 2022; Zhao et al., 2024). However, for children, a very important part of revising a word meaning is learning what building blocks are relevant to begin with, for example, the 'characteristic-to-defining' shift (Keil & Batterman, 1984) in kinship learning. Examining such developments in the feature space would account for a crucial part of slow mapping that our model does not yet capture.

Overall, the model and results reported here contribute to our limited understanding of slow mapping words to meaning. We have shown that behavior ties in with an account of slow mapping in which learners store and change their hypothesis in local and incremental way. In doing so, we have also linked the process of slow mapping to other similar models in higher-order cognition, paving the way for further interdisciplinary work.

Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh,

School of Informatics and School of Philosophy, Psychology & Language Sciences.

References

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, *74*, 35–65.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, *124*(3), 301.

Bramley, N. R., Zhao, B., Quillien, T., & Lucas, C. G. (2023). Local search and the evolution of world models. *Topics in Cognitive Science*.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive psychology*, *96*, 1–25.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Fränken, J.-P., Theodoropoulos, N. C., & Bramley, N. R. (2022). Algorithms of adaptation in inductive inference. *Cognitive Psychology*, *137*, 101506.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, *14*(2), 178–210.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Horst, J. S., & Hout, M. C. (2015). The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior Research Methods*.

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of verbal learning and verbal behavior*, *23*(2), 221–236.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Lakatos, I. (1970). History of science and its rational reconstructions. In *Psa: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1970, pp. 91–136).

Lieder, F., Griffiths, T. L., M. Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic bulletin & review*, *25*, 322–349.

Markham, E., Rabagliati, H., & Bramley, N. R. (2024). Slow mapping words as incremental meaning refinement. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Marr, D. (1982). Vision. *Freeman Co*.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Mollica, F., & Piantadosi, S. T. (2019). Logical word learning: The case of kinship. *Psychonomic Bulletin & Review*, 1–34.

Piantadosi, S. T., & Jacobs, R. A. (2016). Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science*, *25*(1), 54–59.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, *123*(4), 392.

Quine, W. V. O. (1969). *Word and object*. MIT press.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in cognitive sciences*, *20*(12), 883–893.

Sandhofer, C. M., & Doumas, L. A. (2008). Order of presentation effects in learning color categories. *Journal of Cognition and Development*, *9*(2), 194–221.

Suwa, H., & Todo, S. (2010). Markov chain monte carlo method without detailed balance. *Physical review letters*, *105*(12), 120603.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, *66*(1), 126–156.

Tulving, E. (1972). Organization of memory. In (p. 381).

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, *18*(5), 414–420.

Zhao, B., Lucas, C. G., & Bramley, N. R. (2024). A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, *8*(1), 125–136.