OPEN MIND
Discoveries in Cognitive Science

an open access journal

REPORT

Infinite Mixture Chaining: An Efficiency-Based Framework for the Dynamic Construction of Word Meaning

Lei Yu[1] and Yang Xu[1,2]

[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Cognitive Science Program, University of Toronto, Toronto, Canada

Keywords: lexicon, lexical evolution, semantic change, word meaning extension, semantic chaining, probabilistic model, cognitive efficiency

ABSTRACT

The lexicon is an evolving symbolic system that expresses an unbounded set of emerging meanings with a limited vocabulary. As a result, words often extend to new meanings. Decades of research have suggested that word meaning extension is non-arbitrary, and recent work formalizes this process as cognitive models of semantic chaining whereby emerging meanings link to existing ones that are semantically close. Existing approaches have typically focused on a dichotomous formulation of chaining, couched in the exemplar or prototype theories of categorization. However, these accounts yield either memory-intensive or simplistic representations of meaning, while evidence for them is mixed. We propose a unified probabilistic framework, *infinite mixture chaining*, that derives different forms of chaining through the lens of cognitive efficiency. This framework subsumes the existing chaining models as a trade-off between representational accuracy and memory complexity, and it contributes a flexible class of models that supports the dynamic construction of word meaning by automatically forming semantic clusters informed by existing and novel usages. We demonstrate the effectiveness of this framework in reconstructing the historical development of the lexicon across multiple word classes and in different languages, and we also show that it correlates with human judgment of semantic change. Our study offers an efficiency-based view on the cognitive mechanisms of word meaning extension in the evolution of the lexicon.

INTRODUCTION

A primary function of the lexicon is to support the expression of an unbounded set of emerging meanings with a limited vocabulary. As a result, words often take on new meanings. For example, the word *face* in English originally signifying "body part" was extended later to convey meanings such as "facial expression" and "front surface of an object" (Kay et al., 2015). Similarly, the word *store* took on a variety of novel noun arguments including *food, electricity*, and *password* over the past centuries, as illustrated in Figure 1. Word meaning extension is a dynamic process in which words acquire new referents and senses over time, and it is a manifestation of language change which results from a functional need for maximizing communicative expressivity under minimum effort (Blank, 1999; Jespersen, 1959). Existing research has suggested that word meaning extension is non-arbitrary and can be explained partly by

Citation: Yu, L., & Xu, Y. (2025). Infinite Mixture Chaining: An Efficiency-Based Framework for the Dynamic Construction of Word Meaning. *Open Mind: Discoveries in Cognitive Science*, *9*, 1–24. https://doi.org/10.1162/opmi_a_00176

DOI:
https://doi.org/10.1162/opmi_a_00176

Received: 18 April 2024
Accepted: 27 October 2024

Competing Interests: The authors declare no conflict of interests.

Corresponding Author:
Lei Yu
jadeleiyu@cs.toronto.edu

The MIT Press

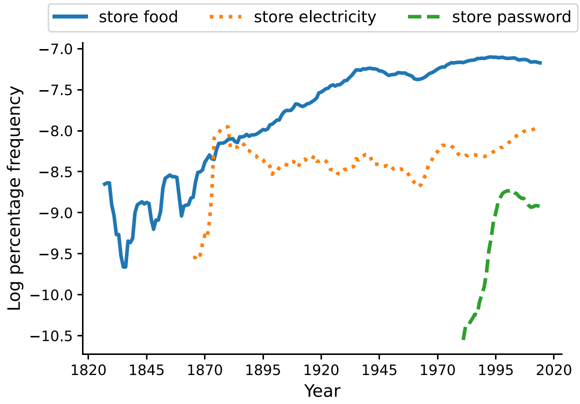


Figure 1. Usage frequencies of the phrases *store food, store electricity*, and *store password* in the past two centuries of English. Data were extracted from Syntactic N-grams historical corpus (Goldberg & Orwant, 2013).

cognitive models of semantic chaining (Grewal & Xu, 2021; Habibi et al., 2020; Hilpert, 2007; Lakoff, 1987; Malt et al., 1999; Ramiro et al., 2018; Sun et al., 2021; Xu et al., 2016; Yu & Xu, 2021). However, there is no unified account of different cognitive models of chaining. We present a general probabilistic framework that derives different forms of semantic chaining through the lens of cognitive efficiency.

Word meaning extension is a form of word meaning change (also known as semantic change), which is a long-standing topic of interest to scholars in historical linguistics (e.g., Bréal, 1897; Traugott & Dasher, 2001) and cognitive linguistics (e.g., Lakoff, 1987). Recent work in cognitive science suggests that word meaning extension is a dominant strategy for lexicalizing emerging meanings in the historical development of the English lexicon, and this process relies in part on semantic chaining, also abbreviated as chaining (Ramiro et al., 2018). Chaining refers to incremental mechanisms of word meaning extension whereby new meanings link to existing ones of a word when they are proximal in semantic space, therefore forming chain-like semantic structures over time (Hilpert, 2007; Lakoff, 1987; Malt et al., 1999; Perek, 2018). Existing studies have tested this incremental view by developing computational models of chaining that account for the historical meaning extension of container names (Sloman et al., 2001; Xu et al., 2016), numeral classifiers (Habibi et al., 2020), adjectives (Grewal & Xu, 2021), verbs (Yu & Xu, 2021), and slang terms (Sun & Xu, 2022; Sun et al., 2021).

These previous studies of semantic chaining typically formulate the models in the tradition of two psychological theories of categorization based on prototypes and exemplars. The prototype theory postulates that each lexical category is represented by a central prototype (Lakoff, 1987; Reed, 1972; Rosch, 1975) and has influenced subsequent cognitive linguists who view chaining as a mechanism for generating radial categories. Another account is the exemplar theory of categorization, which proposes that each category is represented by its set of exemplars stored in memory, and it has motivated a line of computational models like the Generalized Context Model that are commonly used to explain human categorization (Ashby & Alfonso-Reese, 1995; Nosofsky, 1986).

Two outstanding issues emerge from these previous studies. First, they often focused on a dichotomous comparison and assumed that the prototype and exemplar models of categorization are sufficient to capture the cognitive processes of chaining in word meaning extension. Second, which of these two models better explains empirical data has received mixed views in different lexical semantic domains (see Grewal & Xu, 2021; Habibi et al., 2020; Yu & Xu, 2021, but also Geeraerts, 1997; Sun et al., 2021), and therefore there is no unified understanding for the different forms of chaining.

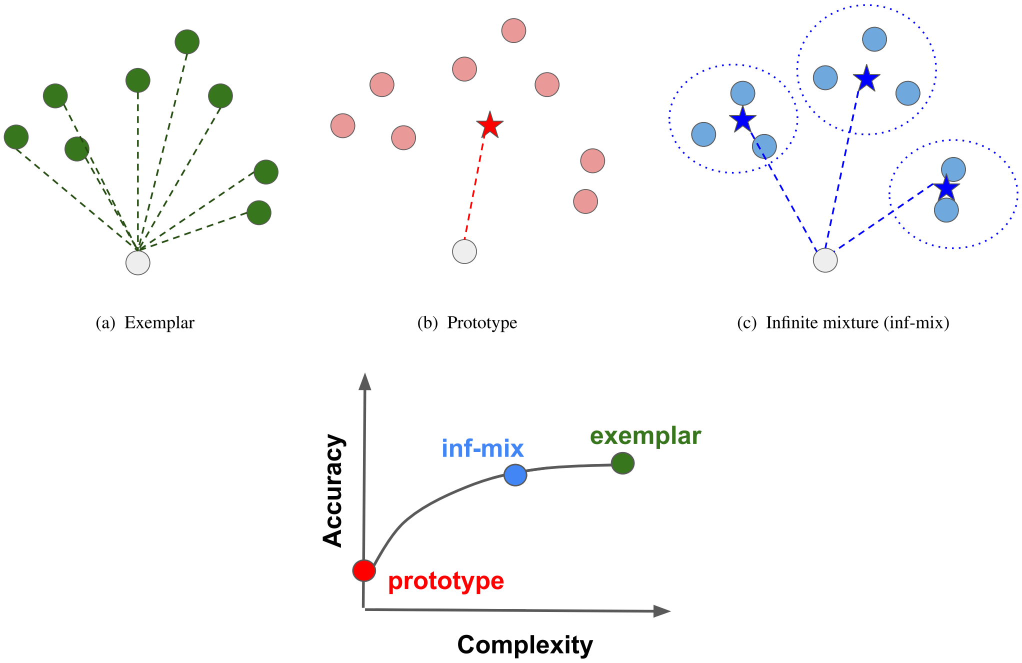
Here we propose a general approach to modeling the dynamic construction of word meaning. Our framework not only subsumes the different computational models of chaining but also offers a new class of flexible models that goes beyond the existing accounts of chaining to support the automatic construction of word meaning through time. In particular, we formulate word meaning extension as the process where a target set of head words expand their collocation classes to pair with a larger array of arguments over time, a view that is consistent with studies on lexical semantic change (Allan & Robinson, 2012; Hilpert, 2012). We evaluate our framework against large-scale historical data in different word classes and languages, as well as human judgment of lexical semantic change.

Theoretical Foundation

Our framework builds on the view that word meanings are structured to support efficient communication (Kemp et al., 2018; Zaslavsky et al., 2018), and that accounts of word meaning extension should take cognitive efficiency into consideration (Ramiro et al., 2018; Xu et al., 2020). Here we define cognitive efficiency as a principled criterion for deriving different formal accounts of semantic chaining, which is based on a tension between two competing constraints that trade off against each other: *representational accuracy* and *memory complexity* (or memory load). Representational accuracy refers to the precision at which a model captures the representation of word meaning, particularly how meaning dynamically changes over time. Memory complexity refers to the cost incurred by a model in meaning representation, particularly the number of stored items required to represent the meanings of words.

Under this view, we propose that the existing accounts of chaining including the prototype and exemplar models can be understood as candidates that fall under the two extremes of this trade-off. At one extreme, the exemplar model offers a highly accurate mental representation of a word's meaning, or lexical category, by storing the past exemplars (i.e., word usages), and it may therefore predict the state of a new item in relation to all the exemplars from memory (see Figure 2a). In this respect, the exemplar model maximizes representational accuracy but at the expense of a high memory load. At the other extreme, the prototype model offers a highly compact representation for a category in terms of a central prototype, and it predicts the state of a new item in relation to that prototype (see Figure 2b). In this respect, the prototype model minimizes memory complexity but at the expense of a simplistic representation, and as a result, the capability of prototype model in predicting or explaining linguistic category extension may be limited compared to exemplar models. The exemplar-prototype dichotomy can thus be interpreted in a unified way as a fundamental tradeoff of cognitive efficiency in word meaning extension: An accurate model tends to demand a high memory load, while a minimum-effort model tends to be more impoverished in representational precision. This efficiency-based framework also offers the possibility to formulate alternative accounts of chaining that go beyond the existing models by near-optimally trading off between the two described constraints of efficiency.

Our proposal is closely related to research in rational human learning and machine learning based on infinite mixtures. Building on this line of work, we model word meaning as an infinite or growing mixture of clusters of usages (see Figure 2c). This modeling scheme can be flexibly adapted to alter the internal structure of a semantic category as it assimilates new items (Alishahi & Stevenson, 2008; Anderson, 1990; Griffiths et al., 2007; Rosseel, 2002; Vanpaemel et al., 2005). In our case, an infinite mixture approach to modeling chaining can potentially



(a) Exemplar (b) Prototype (c) Infinite mixture (inf-mix)

(d) Accuracy-complexity tradeoff

Figure 2. Illustrations of (a–c) models of chaining and (d) how they trade off between representational accuracy and memory complexity in the process of word meaning extension. The exemplar model yields high representational precision by linking a novel item (grey dot) to all existing support items (green dots), so it requires high memory complexity. The prototype model requires low memory by linking the novel item to the prototype (red star), but it tends to be less accurate in representation. The infinite mixture model trades off between accuracy and complexity by constructing a semantic space that groups similar items into a sparse set of clusters (dashed circles), and then linking the novel item to the cluster centroids (blue stars).

capture polysemy (Klein & Murphy, 2001; Li & Joanisse, 2021; Rodd et al., 2012; Tuggy, 1993) and complex structures of word meaning as new items (or usages) emerge over time, therefore moving beyond the exemplar and prototype models which either represent word meanings as a set of independent exemplars or a single prototype. On the other hand, our framework also carries the potential to capture idiosyncratic word usages by storing them as “microclusters” consisting of a single example, therefore offering a “mid-level” structure that lies between schematic and exemplar representations (Dąbrowska, 2004). Similar approaches have been explored in statistical machine learning in the tradition of Dirichlet process (DP) mixture (Allen et al., 2019; Escobar & West, 1995; Ferguson, 1973) which instantiates a trade-off between information loss in model reconstruction of data and complexity in terms of the number of clusters inferred by model (Kulis & Jordan, 2012). Importantly, we demonstrate that the semantic structures inferred by the infinite mixture model can be utilized to predict human judgment of historical semantic change.

COMPUTATIONAL FRAMEWORK

In the following, we first formulate word meaning extension as a temporal prediction problem under the two constraints of cognitive efficiency described. We then show how several existing classes of chaining models and a new class of models can both be derived from this framework, and we specify the semantic space in which these models are operationalized. Code and data for our work are made available in the following repository: https://osf.io/nsrph/?view_only=9c3f9c0abb9b4679ad5cea7dd6ab3a6e.

Problem Formulation Under Efficiency Constraints

We define word meaning extension as a temporal prediction problem following the procedures from existing work on semantic chaining that focused on comparing the exemplar and prototype models (Grewal & Xu, 2021; Habibi et al., 2020; Yu & Xu, 2021).

Given a novel emerging item denoted by n (e.g., this item can be a new concept that emerges over time such as *password*), we want to predict which existing words in the vocabulary can be extended and chosen to describe it. As a concrete example, we first describe our framework in the case of predicting verb meaning extension, and we show later how it can be applied broadly to other word classes not restricted to verbs.

We cast the problem of word meaning extension as probabilistic inference for predicting novel compositional usages of existing verbs and novel noun arguments. Specifically, given a novel noun n such as *password*, we ask which verbs w can be taken as its syntactic predicate to form previously unattested compositions that extend the meaning space of w, e.g., *store*: "to store food" → "to store password". We therefore call the word w as the *head word* (or *head*), and the nouns that form phrases with w the *arguments* of the head. Since a head can take noun arguments under different syntactic roles (e.g., English verbs can take nouns as direct objects, and English adjectives serve as the modifiers of their noun arguments), we also constrain the syntactic relations r in predicting novel phrases. Formally, we consider a head-relation pair (w, r) (e.g., w = *store*, r = *direct object*) as a time-varying category denoted by $\mathcal{S}^{(t)}_{w,r}$ that consists of all existing syntactic noun arguments under relation r up to time t. The temporal inference problem is then equivalent to predicting the probability of any query noun n_q to emerge in that category at a future time. We focus on predicting pairings of existing heads with query nouns that have not yet appeared as arguments for given heads – for instance, the category "*store (direct object)*" may have been attested include the noun *food* up to time t, and can be predicted by a model to extend toward predicating new technological terms such as *information* or *password* later.

Formally, given an emergent query noun n_q at time t, and a list of head word forms w with existing noun arguments $\mathcal{S}^{(t)}_{w,r}$ up to t, our framework models the process of semantic chaining in the **head prediction problem** – i.e., inferring which heads will be appropriate predicates for n_q under syntactic relation r at time $t + \Delta$, where Δ is an incremental time step. The probability of a head w being a predicate of n_q via relation r is defined as follows:

$$p\left(w, r|n_q\right)^{t+\Delta} = p(n_q|\mathcal{S}^t_{w,r}) \propto \mathrm{sim}(n_q, \mathcal{S}^t_{w,r}) \tag{1}$$

Our framework can be applied similarly to the **argument prediction problem** asking which novel query nouns will most likely emerge into a head's referential range in the near future. We achieve this by modeling the conditional probability $p(n_q| w, r)$ of n_q being added as a new syntactic argument of w. We take a probabilistic Bayesian approach by computing $p(n_q| w, r)$ as a posterior distribution with a frequency-based prior over heads:

$$p\left(n_q|w, r\right)^{t+\Delta} \propto p_0(w, r)^t p\left(w, r|n_q\right)^{t+\Delta} \propto N(w, r)^t \mathrm{sim}(n_q, \mathcal{S}^t_{w,r}) \tag{2}$$

where $N(w, r)^t$ is the observed frequency count of w being a syntactic head of some arguments via relation r up to time t.

Importantly, $\mathrm{sim}(n_q, \mathcal{S}^t_{w,r})$ in both of the above equations is a yet-to-be-specified function (i.e., implementing different ways of chaining) that measures the semantic similarity between the query noun and current meaning of the head-relation category $\mathcal{S}_{w,r}$ at time t. To compute this similarity, we quantify the semantic proximity between n_q and the existing set of noun arguments of $\mathcal{S}_{w,r}$ (i.e., category exemplars). We refer to this set of nouns as the *support set* (denoted by $n_s \in \mathcal{S}_{w,r}$). We assume that the semantic similarity between a query noun and a support set can be captured by the semantic distances between the query and a set of cluster centroids inferred among the support nouns which we denote as $\mathcal{M}_{w,r}$.

$$\mathrm{sim}(n_q, \mathcal{S}^t_{w,r}) = \mathrm{sim}(n_q, \mathcal{M}^t_{w,r}) = \mathrm{sim}(n_q, \{\mu^t_{w,r,k}\}_{k=1}^{K^t_{w,r}}) \tag{3}$$

Here $\mathcal{M}^t_{w,r} = \left\{\mu^t_{w,r,1}, \mu^t_{w,r,2}, \ldots\right\}$ is a set of $K^t_{w,r}$ cluster centroids for support set $\mathcal{S}^t_{w,r}$. In the next section, we show that exemplar chaining is equivalent to the case where each support noun (or exemplar) forms its own cluster; prototype chaining is the case where all support nouns are represented as a single cluster; and infinite mixture chaining sits in between these two extremes.

We quantify every noun n at a given time using distributed semantic representation $\phi(n)^t$ in a high dimensional space that changes over time (details specified in the later section on diachronic semantic space). Following the psychological literature (Nosofsky, 1986), we define semantic similarity as the mean negative exponential Euclidean distance between the query noun and the cluster centroids of a word-relation category:

$$\mathrm{sim}(n_q, \mathcal{M}^t_{w,r}) = \frac{1}{K^t_{w,r}} \sum_{k=1}^{K^t_{w,r}} \exp\left(-\frac{\left\|\phi(n_q)^t - \mu^t_{w,r,k}\right\|^2}{\beta}\right) \tag{4}$$

where we follow the Generalized Context Model and its variants (Kruschke, 2008; Maddox & Ashby, 1993; Nosofsky, 1986) by adding a sensitivity parameter β that controls the rate at which similarity decreases with semantic distance. We allow the number of cluster centroids to flexibly vary over time (as a head word encounters new nouns), which is inferred and updated based on the internal semantic structure of a head-relation category instantiated in terms of its support nouns. In particular, the semantic clusters inferred within a category are expected to optimize the following trade-off between two constraints of efficiency, following work on infinite mixtures from machine learning (Kulis & Jordan, 2012):

$$\mathcal{M}^t_{v,r} = \mathrm{argmin}_{\mathcal{M}} \sum_{k}^{K^t_{v,r}} \sum_{n_s \in \mathcal{S}^t_{v,r}} \left\|\phi(n_s)^t - \mu^t_k\right\|^2 + \lambda K^t_{v,r} \tag{5}$$

The first term on the right of Equation 5 is known as the information loss, which quantifies how accurately a set of cluster centroids can represent the full set of support nouns (e.g., in the exemplar model, representational accuracy is near ceiling because each exemplar is in its own cluster). The second term measures the memory complexity for storing cluster centroids (e.g., in the prototype model, memory complexity for a given word is 1, which is the theoretical minimum if we wish to avoid zero representational accuracy). A single parameter λ controls the relative weighting between the two constraints. Intuitively, models with higher values of λ would favor a more parsimonious approach of chaining by inferring as few clusters as possible (with prototype model at the extreme), while models with smaller values of λ would store as many clusters as possible to minimize information loss (with exemplar model at the extreme).

Our formulation of the efficiency tradeoff is also related to the information bottleneck theory of efficient communication, which assumes that word meanings are organized under the tradeoff between reconstruction accuracy and complexity (Tishby et al., 2000; Zaslavsky et al., 2018). However, a crucial distinction here is that our focus is on model inference of newly emerging meanings for individual words rather than the construction of a semantic system where word meanings are held static.

Classes of Chaining Model

The efficiency formulation in Equation 5 helps derive several classes of chaining model from the literature and anew, and we show that our framework subsumes these classes under a broader spectrum of models.

Exemplar-Based Models. In the case where the tradeoff parameter λ → 0, the model ignores the memory constraint and stores every support noun argument n_s as a single cluster to achieve zero information loss. The inf-mix model therefore boils down to the exemplar model of chaining:

$$p\left(n_q|w,r\right)^{t+\Delta} \propto \frac{1}{|S^t_{w,r}|}\sum_{n_s\in S^t_{w,r}} \exp\left(-\frac{\left\|\phi\left(n_q\right)^t - \phi\left(n_s\right)^t\right\|^2}{\beta}\right) \tag{6}$$

The literature has also suggested that a variant of the exemplar model, particularly 1-nearest-neighbor (1nn) chaining, has been effective in predicting emergent word senses (Ramiro et al., 2018). If we adjust the inference procedure by considering only one support noun closest to the query noun (in semantic space) instead of all the support nouns, we can easily derive the 1nn chaining model:

$$p\left(n_q|v,r\right)^{t+\Delta} \propto \text{argmax}_{n_s\in S^t_{v,r}} \exp\left(-\frac{\left\|\phi\left(n_q\right)^t - \phi\left(n_s\right)^t\right\|^2}{\beta}\right) \tag{7}$$

Prototype Model. If $\lambda \to \infty$, the model yields a minimal memory cost by storing only a single cluster centroid (or the prototype) for each category, and it therefore converges to the prototype model:[1]

$$p\left(n_q|w,r\right)^{t+\Delta} \propto \exp\left(-\frac{\left\|\phi\left(n_q\right)^t - \mu^t_{w,r}\right\|^2}{\beta}\right) \tag{8}$$

Here $\mu^t_{w,r} = \frac{1}{|S^t_{w,r}|}\sum_{n_s\in S^t_{w,r}}\phi(n_s)$ is the mean embedding of all nouns in a support set.

Infinite Mixture Model (inf-mix). In the intermediate cases where $0 < \lambda < \infty$, the number of clusters lies between 1 and the support set size $|S^t_{w,r}|$ and can be inferred using a deterministic algorithm called DP-Means (Kulis & Jordan, 2012). This is a nonparametric variation of the well-known K-means clustering algorithm in unsupervised learning (Hartigan & Wong, 1979). The centroids $\mathcal{M}^t_{w,r}$ would then be the mean vector representation of the support arguments within each cluster. Figure 2 illustrates the different classes of chaining model in the computation of $p(n_q|w,\ r)$. Theoretically, it can be shown that the infinite mixture chaining model is equivalent to the asymptotic case of a Dirichlet Process Gaussian Mixture Model (DPGMM) (Görür & Rasmussen, 2010) with the variance parameter of the Gaussian likelihood function shrunk toward 0 (Kulis & Jordan, 2012). However, in a fully Bayesian DPGMM, the

[1] Precisely, the inf-mix model will become the prototype model as long as λ is greater than the maximum pairwise Euclidean distance between any two support noun embeddings.

mixture centroids μ_k become latent variables and need to be inferred via posterior sampling, which requires storing and repeatedly using all support noun arguments. This is computationally prohibitive for common head-relation classes consisting of hundreds or even thousands of noun arguments. Our framework bypasses these issues of DPGMM and is more computationally efficient.

Semantic Space

The chaining models described need to be operationalized in a time-varying semantic space so that information about future head-argument phrases should be minimally smuggled into prediction at current time points. We use Word2Vec-based representations commonly used in natural language processing for distributed semantics (Mikolov et al., 2013). Note that word co-occurrence distributions are constantly changing and therefore the semantic space needs to be updated to capture information only up to time *t*. For this reason, we use the 300-d HistWords pre-trained diachronic embeddings (Hamilton et al., 2016), where the embedding for each noun at decade *t* is based solely on its co-occurrence statistics from the current decade, while the future co-occurrences are not embedded. Other studies have explored multimodal representations of word meaning beyond textual data (Brochhagen et al., 2023; De Deyne et al., 2021; Yu & Xu, 2021), which can provide alternative semantic representations.

DATA

We evaluate our infinite mixture chaining framework on reconstructing historical extension in three classes of words, derived from three separate datasets building on the existing literature of semantic chaining: 1) English verb phrases consisting of head verbs and noun objects (cf. Yu & Xu, 2021), 2) English adjective phrases of head adjectives and modified nouns (cf. Grewal & Xu, 2021), and 3) Chinese numeral classifiers and their measured nouns (cf. Habibi et al., 2020). Table 1 shows sample entries from these datasets which we describe in turn.

Historical Data of English Verb Phrases

Building on the study of Yu and Xu (2021), we collected a large dataset of historical head-argument compositions derived from the Google Syntactic N-grams (GSN) English corpus, where the noun argument can be either the direct object of the verb (e.g., store the *password*) or can be an indirect prepositional object (e.g., store *in the computer*). In particular, we collected verb-noun-relation triples $(n, v, r)^t$ that co-occur in the ENGALL subcorpus of GSN from 1850 to 2000. We focused on working with common usages and pruned rare cases under the following criteria: 1) all noun arguments are extracted from a large vocabulary of words with top-10,000 noun counts (with POS tag as noun) in GSN over the 150-year period; 2) all verbs

Table 1. Sample entries from Google Syntactic N-grams including head-relation pairs, support and query nouns, and timestamps.

	Head-relation pair			
Decade	Head word	Syntactic relation	Support noun	Query noun
1900	drive (verb)	direct object	horse, wheel, cart	car, van
1950	work (verb)	prepositional object via *as*	mechanic, carpenter, scientist	astronaut, programmer
1980	healthy (adjective)	modified objec	food, diet, life	vegan, finance
2000	次(cì) (Chinese classifier)	modified object	资助(funding), 就业(employment), 发言(speech)	公投(referendum)

should have at least 20,000 counts in GSN. To support feasible computations, we consider the top-20 most common syntactic relations in GSN between head verbs and noun arguments. We binned the raw co-occurrence counts by decade $\Delta = 10$. At each decade, we define emerging noun arguments for a given verb-relation category (v, r) if their number of co-occurrences with (v, r) up to time t falls below a threshold θ_q, while the number of co-occurrences with (v, r) up to time $t + \Delta$ is above θ_q (i.e., an emergent usage that conventionalizes over time, as opposed to a spontaneous usage). We define support nouns as those that co-occurred with (v, r) for more than θ_s times before t. We found that $\theta_q = 10$ and $\theta_s = 100$ are reasonable choices. This preprocessing pipeline yielded a total of 8,897 verb-relation categories of noun arguments over 15 decades, where each category has at least 1 novel query noun and 10 existing support nouns in each decade.

Historical Data of English Adjective Phrases

Analogous to verb extension and building on the study of Grewal and Xu (2021), we extracted historical compositions of English adjective modifiers and their noun arguments from the ENGALL subcorpus of GSN from 1850 to 2000. We also pruned rare usages by keeping only the same set of top-10,000 most frequent noun types as in the verb phrase dataset, and removing phrases whose head adjective has less than 20,000 counts in GSN. At each decade, the criteria of deciding novel emergent noun arguments for an adjective is the same as those described in the verb phrase collection pipeline. We finally obtained a total of 2,037 adjective categories of noun arguments over 15 decades, where each category has at least 1 novel query noun and 10 existing support nouns in each decade.

Historical Data of Chinese Numeral Classifiers

Building on the study of Habibi et al. (2020), we also apply inf-mix chaining to reconstruct meaning extension of linguistic categories beyond English. Specifically, we consider how Chinese numeral classifiers have been applied to modify novel nouns in the 20th century. Chinese classifiers are obligatory grammatical classes that are used between a noun and a numeral term describing its quantity, e.g., one [个(gè)] person or two [份(fèn)] documents. We made use of the data of historical linguistic category extension in Habibi et al. (2020), which includes a comprehensive list of 8,371 (Chinese classifier, noun) pairs over the period 1940–2003. We follow Habibi et al. (2020) by representing each Chinese noun as the pretrained word2vec contemporary embedding of its English translation to prevent information smuggling. Different from Habibi et al. (2020) that predicted novel usages on a yearly basis, we binned the noun-classifier pairs by $\Delta = 5$ based on their time of emergence to ensure that between two consecutive evaluation time points, there is a sufficient number of newly introduced arguments and a set of significantly different inferred semantic centroids by inf-mix models.

RESULTS

In this section, we first show via a simulation study that inf-mix chaining offers better account for the extension process of categories with various structures. We then demonstrate in three case studies that cognitive efficiency can derive new chaining models that balance the tradeoff between predictiveness of new meanings and memory complexity. We finally show that the semantic representations derived from inf-mix chaining is psychologically grounded in that they can capture human judgment of diachronic word meaning change.

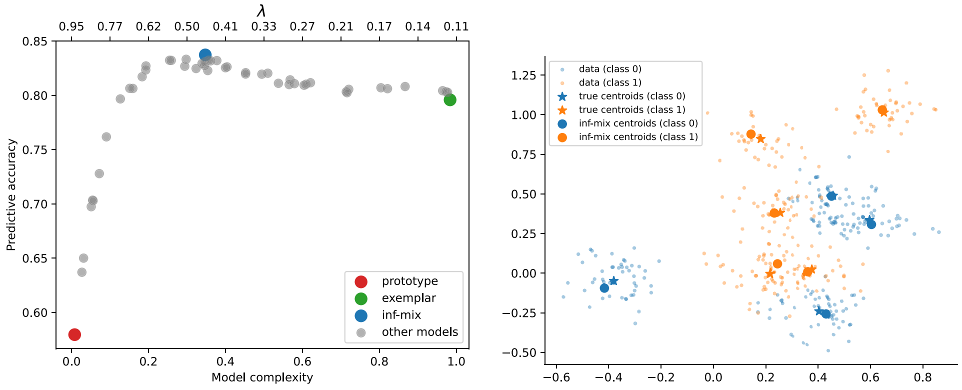
Infinite Mixture Chaining on Simulated Data

We first compare inf-mix and existing models of chaining using simulated data of category extension in a continuous two-dimensional space. To do so, we generate $N = 500$ data points for each of two competing category via Gaussian Mixture models with randomly sampled components. We then present a chaining model with 80% of generated points from each category as the training set to infer category centroids and use the model to predict category labels for the remaining 20% as the test set. We test a series of 50 inf-mix chaining models whose inferred number of semantic centroid ranges from 1 (prototype) up to 500 (exemplar) with a step-size of 10. In each simulation, the number of Gaussian component for each category is randomly sampled between 1 and 10, and the mean of each Gaussian distribution is randomly drawn from a 2-D uniform distribution in $[-1, 1] \times [-1, 1]$. We keep all Gaussian distributions isotropic with a variance of 0.1 in both dimensions.

Figure 3a illustrates the trade-off between mean model predictive accuracy on test set and normalized model complexity measured by the ratio between the number of inferred clusters and the category size. We found that a range of inf-mix models with moderate complexities best balance the trade-off by offering the highest predictive accuracy while maintaining a relatively low memory cost. In contrast, the exemplar model is less accurate and more memory-intensive, while the prototype model, despite having the lowest memory cost, yields the lowest predictive accuracy that is only slightly above chance. Figure 3b shows the generated data points of two Gaussian mixture categories with 4 and 5 components respectively, together with the inferred cluster centroids by the inf-mix model of highest test accuracy. In this case, the inf-mix model almost perfectly recovers the ground-truth centroids for both categories. These simulation results suggest that for artificial categories with various structures, inf-mix chaining has the potential to offer the most cognitively efficient account of category extension.

Case Study 1: English Verb-Noun Compositions

We next evaluated empirically different classes of chaining models on both head and argument prediction problems under variation of the trade-off parameter λ on emerging verb-noun compositions for the historical period 1850 to 2000. At every decade, for argument prediction



(a) (b)

Figure 3. (a) Predictive accuracy and memory complexity (ratio between inferred cluster number and category size) on 2-D simulated data. (b) Example generated data points and inferred cluster centroids by inf-mix for two competing categories with multiple Gaussian mixture components.

problem, we randomly sample for each verb-relation pair (v, r) 100 candidate noun arguments (with exactly one of them being a ground-truth novel argument n_q) from the vocabulary of top-10,000 nouns in GSN, and we then compute the percentage of cases where each chaining model predicts the true n_q over the random noun set as a more appropriate argument. Similarly, for the head prediction problem, we randomly sample for each query noun 100 candidate verb-relation pairs (with exactly one of them being a ground-truth head of the query noun) from the vocabulary of top-5,000 verb-relations in GSN. This procedure allows us to assess the degree to which each class of chaining model can successfully predict novel verb-noun compositions incrementally through time, and how they fare in the accuracy-complexity trade-off. To pursue best model predictability of word meaning extension, we apply stochastic gradient descent to tune the sensitivity parameter β in the semantic similarity functions of each chaining model to maximize their average predicted probability $p(n_q|w, r)^{t+\Delta}$ over all ground-truth (novel noun argument, head word, relation) triples in the dataset.

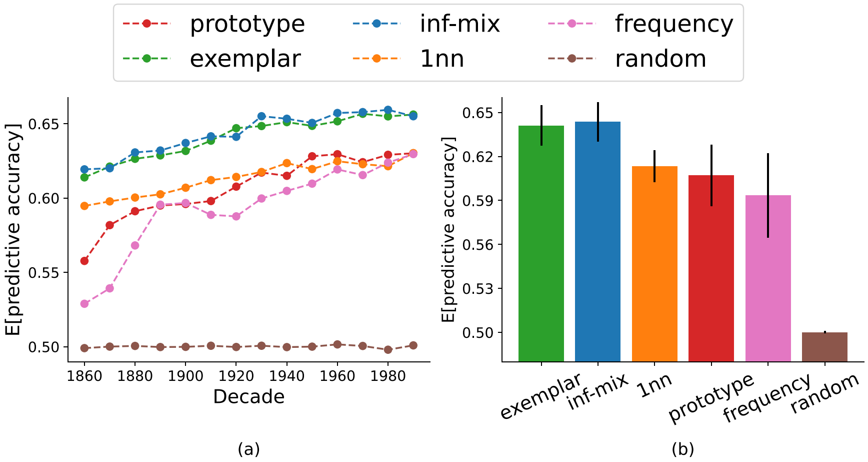
For infinite mixture models with $0 < \lambda < \infty$, we implemented the DP-means clustering algorithm introduced in Kulis and Jordan (2012) to assign a categorical cluster label for every noun within the support set of each verb-relation pair, and take the mean word embeddings of support nouns in each inferred cluster as centroid to compute the likelihood function $p(n_q| v, r)$. Since Euclidean-distance-based clustering methods such as DP-means tend to degenerate on high dimensional data (due to the curse of dimensionality), we instead perform DP-means on a 30-dimensional subspace of the HistWords embeddings projected by principal components analysis (PCA). We found that this reduced subspace preserves well the relative distances between word pairs (explaining over 80% of variance from the original 300-dimensional data) and yields reasonable clustering results. During prediction, we use the full HistWords embeddings by computing centroids using clustering labels computed on the PCA subspace.

We found that when $\lambda = 0.24$ and $\beta = 0.45$, the inf-mix model yields most well-defined clustering overall measured by the standard Silhouette score for unsupervised clustering.[2] We therefore evaluate this model on its predictive accuracy by-decade and aggregate predictive accuracy, along with the other competing models. We also consider two baseline models: a frequency baseline that always favors the noun with the highest usage frequency in GSN up to the decade in question, and a random baseline.

Figure 4 summarizes the results for the main problem of head prediction. We observe that among all the models examined, the infinite mixture and exemplar models yielded near-equivalent predictive accuracy and are superior than the alternatives. The prototype and 1nn models perform better than the two baselines, but they are much worse than the top 2 models. We observed similar result patterns in the argument prediction problem as well (see Figure A1 in Appendix). These initial results show that the infinite mixture model is on par with the exemplar model in predicting historical verb extension, the latter being the better performing model as reported in recent work of chaining (Grewal & Xu, 2021; Habibi et al., 2020; Yu & Xu, 2021).

To assess the efficiency of different chaining models, we computed the expected predictive accuracy from the average predictive percentages over all (v, r, n_q) triples in the dataset. We also measured memory complexity by computing the expected number of cluster centroids inferred for every set of support noun arguments at each decade. We focus on comparing the

[2] We took the averaged Silhouette score over clustering of all support sets across all decades, and found that the inf-mix model with $\lambda = 0.24$ yields the highest mean Silhouette score.



(a)

(b)

Figure 4. Model accuracy of head verb prediction through time (left panel) and in aggregate (right panel). The infinite mixture model has $\lambda = 0.24$. Error bars represent the standard deviations of accuracy across decades.

infinite mixture model with the two most representative models of chaining, prototype and exemplar. We also incrementally vary λ to assess a large set of other alternative classes of chaining beyond the three target models. Figure 5 shows the result for head verb prediction which indicate that 1) by sweeping λ from 0 toward ∞ (in this case $\lambda \geq 0.5$ suffices), the predictive accuracy drops only slightly from the exemplar model to the infinite mixture model ($\lambda = 0.24$) but substantially to the prototype model—this finding confirms with our previous analysis, that the infinite mixture model predicts on par with the exemplar model; and 2) the marginal gain on accuracy of the exemplar model comes at a high cost in memory

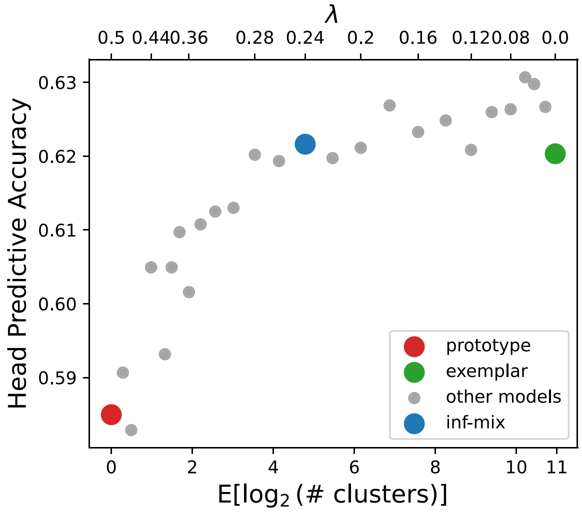


Figure 5. Accuracy of head prediction problem and memory complexity (mean number of clusters per word used in prediction) on English historical verb-noun composition dataset. The gray dots show the spectrum of infinite mixture models under different λ values. The left end (red dot) of the x-axis corresponds to the prototype model with $\lambda \geq 0.5$. The right end (green dot) of the x-axis corresponds to the exemplar model with $\lambda = 0$. The blue dot corresponds to the infinite mixture model with inferred optimal λ value.

complexity: compared to the infinite-mixture model, it requires over 2^5-fold more storage of cluster centroids to achieve a gain of <0.01 in predictive accuracy. We observed a similar trade-off curve for the argument noun prediction problem as well (see Figure A2 in Appendix). Overall, the infinite mixture model achieves a better balance between precision and memory.

We interpret the semantic clusters learned by the infinite mixture model using verb category *store in dobj* as an example. Figure 6 illustrates its meaning space spanned by support nouns in 1900s and 1980s respectively, projected on a 2D plane using the t-distributed Stochastic Neighbor Embedding (van der Maaten & Hinton, 2008). The model identifies 4 clusters of semantically related *store*-able nouns in 1900s and 7 clusters in 1980s, most representative noun arguments for which are shown in the legends of Figure 6. To track how these the meaning clusters change over time, we mark a pair of clusters across the two decades with the same color if they share the highest number of overlapping support arguments. For instance, the cluster with arguments *bean, honey, meat* in 1980s is colored in blue, since it shares the most support nouns with the *corn, flour, wheat* cluster at 1900s. The three clusters in 1980s with distinct colors (marked in olive, cyan and black) can be considered as novel senses that the verb category acquired during the 20th century. We found that the infinite mixture model not only infers consistent noun clusters across time by adding semantically related novel nouns to the existing clusters (e.g., assigning words like *key, data* to the red cluster denoting abstract concepts related to knowledge and mind), but also detects novel word senses by growing clusters that contain those emerging concepts (e.g., the olive cluster of biology terms, and the black cluster that contains information technology terms).

Case Study 2: English Adjective-Noun Compositions

Similar to the previous case study on English verb phrases, we evaluated chaining models trade-off parameter $\lambda \in [0, 0.6]$ on predicting emerging noun arguments for English adjectives between 1850 and 2000. Each trial of head prediction consists of a true attested adjective and 99 randomly sampled adjectives that are never paired with the target noun, and each trial of argument prediction consists of a true attested argument noun and 99 randomly sampled

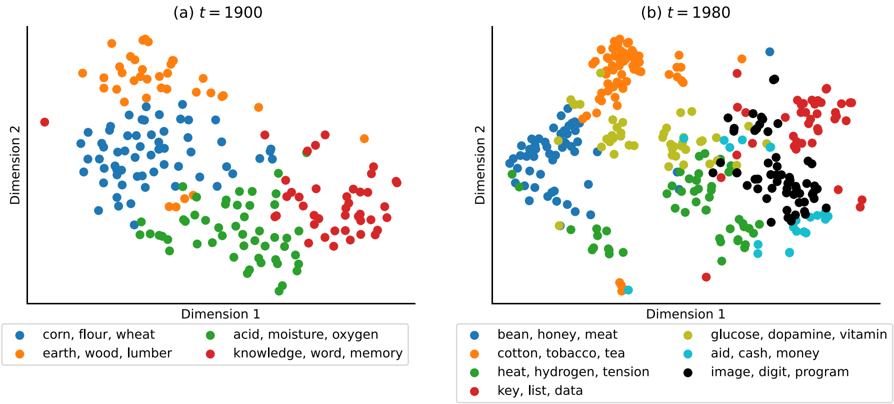


Figure 6. Low-dimensional visualizations of historical meaning extension for the verb frame *store in (noun)* from 1900s (left) to 1980s (right) via t-SNE projection. The dots correspond to word embeddings of noun arguments grouped in clusters inferred by the infinite mixture chaining model. Legends show 3 representative nouns closest to their cluster centroids for each cluster.

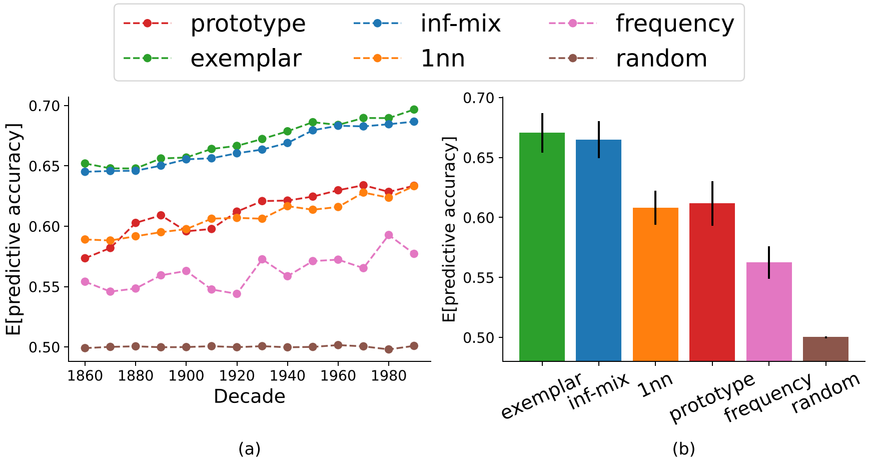


Figure 7. Model accuracy of head adjective prediction through time (left panel) and in aggregate (right panel). The infinite mixture model has $\lambda = 0.14$. Error bars represent the standard deviations of accuracy across decades.

nouns that are never paired with the target adjective. All chaining models are again evaluated on a PCA-reduced 30-dimensional HistWord historical semantic space, in which we also performed DP-Means clustering for inf-mix models and found that $\lambda = 0.14$ and $\beta = 1.50$ yielded the most coherent clustering results (measured by the Silhouette score).

Figure 7 summarizes the predictive accuracy for all chaining models on the main task of head prediction. We again observe that among all the models examined, the infinite mixture and exemplar models yielded near-equivalent predictive accuracy and are superior than the alternatives. Figure 8 summarizes the memory complexity and predictive accuracy on head adjective prediction task for all models. Again, the most coherent infinite mixture model with $\lambda = 0.14$ and exemplar models yielded nearly identical performance, whereas the latter requires approximately four times more memory space to operate. and as λ goes above the critical value of 0.14, the performance of models of decreasing memory cost drops

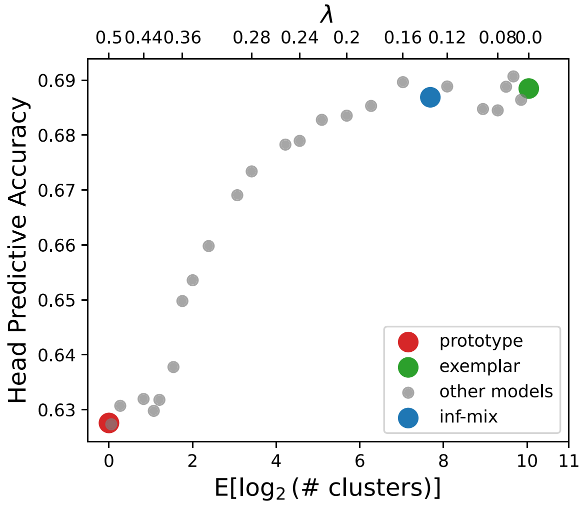


Figure 8. Accuracy of head prediction problem and memory complexity (mean number of clusters per word used in prediction) on English historical adjective-noun composition dataset.

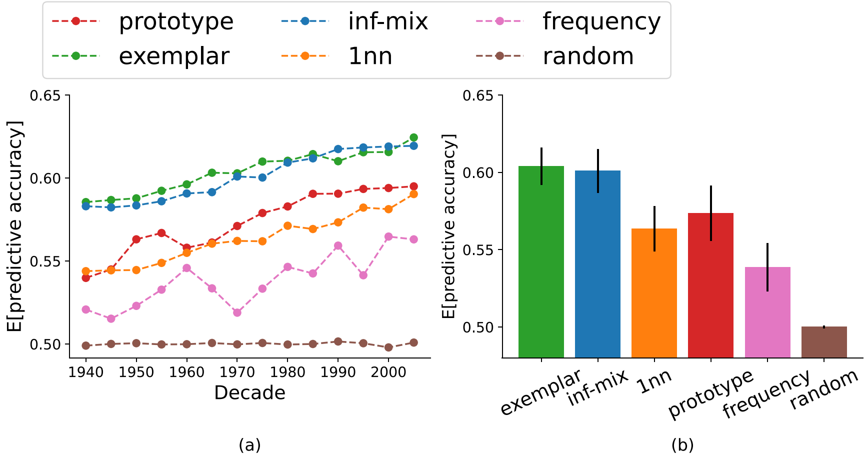


Figure 9. Model accuracy in head classifier prediction through time (left panel) and in aggregate (right panel). The infinite mixture model has $\lambda = 3.0$. Error bars represent the standard deviations of accuracy across decades.

significantly. We observed similar result patterns in the argument prediction problem as well (see Figure A3 and A4 in Appendix). These results conform with our finding in case 1 and suggest that the inf-mix chaining model offers the best trade-off between precision and memory in accounting for historical semantic change of English adjectives.

Case Study 3: Chinese Classifier-Noun Compositions

Figure 9 summarizes the expected head prediction problem accuracy over all (classifier, novel noun) pairs in the Chinese numeral classifier dataset for all chaining models. Similar to the previous two case studies, the infinite mixture and exemplar models are on par with each other while outperforming all other models. Figure 10 summarizes the memory complexity and

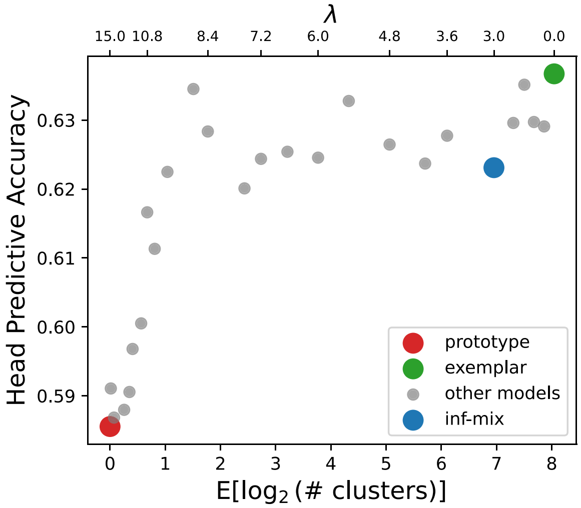


Figure 10. Accuracy of head prediction problem and memory complexity (mean number of clusters per word used in prediction) on Chinese historical noun-classifier composition dataset.

predictive accuracy on the head classifier prediction task. Due to the different geometric properties of the contemporary word2vec embeddings, we found that in this case the inf-mix model converges to the prototype model at a much larger $\lambda \approx 15.0$ compared to thresholds in the previous two experiments, and a sensitivity parameter $\beta = 0.96$. Moreover, we observed that the optimal inf-mix model of the most coherent semantic clusters is more memory-intensive than the previous two optimal inf-mix models (with an optimal $\lambda \approx 3.0$), suggesting that the meaning of Chinese classifiers is much more "polysemous" compared to English verbs and adjectives, as the noun arguments of Chinese classifiers cannot be summarizes using a small set of semantic classes. However, despite the more complex nature of classifier meaning, the performance of the inf-mix model again matches the best exemplar model with about 50% less memory requirement. We observed similar result patterns in the argument prediction problem as well (see Figure A5 and A6 in Appendix).

Taken together, the three case studies suggests that inf-mix offers a novel view and extends existing models of word meaning extension, by incorporating cognitive efficiency into semantic chaining to quantify the trade-off between predictive precision and memory complexity.

Case Study 4: Human Judgment of Lexical Semantic Change

Finally, we perform a quantitative analysis to assess whether the inferred semantic centroids by inf-mix chaining model can be applied to explain people's judgment of diachronic lexical semantic change (LSC) in English. We choose the subtask 2 of SemEval-2020 Task 1 (Schlechtweg et al., 2020) as our evaluation dataset, where an unsupervised model is asked to rank a set of target words according to their degree of lexical semantic change between corpora C_1 and C_2 from two different time periods. For the case of English, C_1 and C_2 are the subsets of the Clean Corpus of Historical American English (CCOHA) (Alatrash et al., 2020) that spans time periods 1810–1860 (C_1) and 1960–2010 (C_2) respectively, and there are 37 target word types (33 nouns and 4 verbs) for evaluation.

We construct an inf-mix model for predicting semantic change in the following way: for each target word w, we first use a deep contextualized neural language model named BERT (Devlin et al., 2019) to encode each usage sentence of w into a high-dimensional sentence embedding space. We then run the DP-Means clustering algorithm on two sets of usage sentence embeddings in C_1, C_2 respectively to obtain two groups of induced centroid embeddings $H_w^{(1)}, H_w^{(2)}$. The degree of semantic change of w between the two time periods of study can then be quantified as the mean pairwise cosine distance between $H_w^{(1)}, H_w^{(2)}$:

$$s_w(t_1, t_2) = \frac{1}{|H_w^{(1)}| \cdot |H_w^{(2)}|} \sum_{h_1 \in H_w^{(1)}, h_2 \in H_w^{(2)}} \frac{\langle h_1, h_2 \rangle}{\|h_1\|^2 \|h_2\|^2} \tag{9}$$

We evaluate the models by computing the Spearman's ρ correlation score between predicted semantic change scores and the gold-standard results by human annotators. Similar to the previous three case studies, we varied the trade-off parameter λ of the inf-mix model from 0 up to a sufficiently large value of 0.7 with a step size of 0.05, and take the one with highest silhouette clustering score as the optimal inf-mix model. The case of $\lambda = 0$ and $\lambda = 0.7$ again correspond to a prototype-based and an exemplar-based LSC model and can be taken as two baselines.

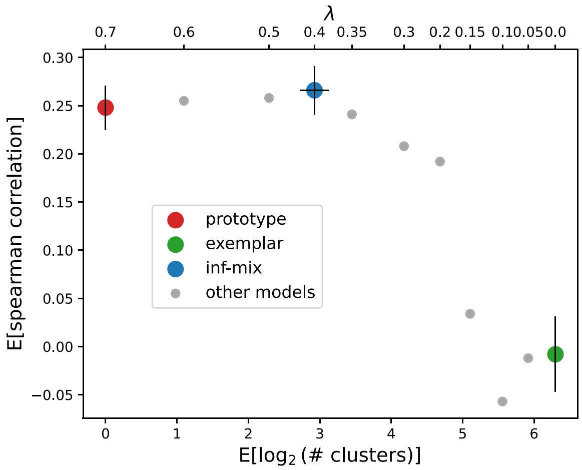


Figure 11. Correlations between inf-mix model predictions and human annotated degree of semantic change for 37 English words between two historical time periods. The gray dots show the spectrum of infinite mixture models under different λ values and varying numbers of inferred semantic centroids (memory complexity). The left end (red dot) of the x-axis corresponds to the prototype model with $\lambda \geq 0.7$. The right end (green dot) of the x-axis corresponds to the exemplar model with $\lambda = 0$. The blue dot corresponds to the infinite mixture model with inferred optimal $\lambda = 0.15$.

Figure 11 shows correlation scores for a series of inf-mix models. We observe that the inf-mix model with highest clustering quality (measured by silhouette score) coincides with the inf-mix model ($\lambda = 0.15$) of highest correlation between human LSC judgements ($\rho = 0.266$, $p < 10^{-11}$) compared to the prototype ($\rho = 0.245$, $p < 10^{-7}$) and the exemplar ($\rho = -0.013$, $p < 10^{-6}$) models. Notably, the exemplar model that yields best predictive accuracy in the previous three case studies now performs much worse compared to models of lower memory costs, suggesting that semantic abstraction plays a more essential role than individual usage memorization in modeling diachronic lexical semantic change. On the other hand, the inf-mix model yields either the best or a near-optimal performance across all four studies, suggesting the important role of cognitive efficiency in explaining word meaning change over time. Table 2 shows several example lemmas on which the inf-mix model yields the best predictive accuracy in LSC. Though not perfect, we found that the inf-mix model is better than the alternative models in capturing prominent word meaning changes (e.g. *tip*), and words with relatively stable meaning (e.g. *chairman*).

Table 2. Sample English lemmas in SemEval-2020 Task 1 with ground-truth and model predicted scores of lexical semantic change (LSC) between the time periods of 1810–1860 and 1960–2010.

Lemma	Ground-truth LSC	Prototype predicted LSC	Exemplar predicted LSC	Inf-mix predicted LSC
chairman (noun)	0	−0.012	0.133	0.008
player (noun)	0.274	0.557	0.182	0.306
tip (verb)	0.679	−0.197	−0.281	0.223

DISCUSSION AND CONCLUSION

We presented an efficiency-based computational framework of semantic chaining for modeling the historical extension of word meaning. Our framework provides a synthesis of existing approaches and suggests that different forms of chaining can be understood as a trade-off between model representational accuracy and memory complexity. Our study moves away from the typical focus on a comparative analysis of different chaining models but instead develops a unified view toward interpreting the diverse kinds of chaining in word meaning extension. In particular, we showed that the most commonly described chaining models based on the exemplar and prototype theories can be interpreted as two extremes of a trade-off for cognitive efficiency.

Our work extends beyond existing work on chaining by proposing a new class of models that supports flexible growth of meaning clusters based on historical and emerging word usages. We showed that the infinite mixture chaining model is on par with the exemplar model and performs better than the prototype model in reconstructing the historically emerged noun-argument pairings with English verbs and adjectives, and numeral classifiers in Mandarin Chinese. We also showed that the same model yields a substantially more compact representation for the internal structures of word meaning compared to the exemplar model, and therefore near-optimally trading off model accuracy with complexity.

Our framework has several limitations. Firstly, we considered a simplified problem of word meaning extension by predicting how words should pair up with previously unattested concepts as they emerged over time. However, word meaning may change without involving novel compositions. For instance, the phrase *to save my key* used to refer to “keeping a physical key”, but the same phrase took on the novel meaning “to save a string that gives access to retrieve information” without requiring any change in its compositional form. Additionally, word meaning extension may occur in a highly contextual setting that is not necessarily reflected in novel argument pairing. For example, saying “that person is sick” can be interpreted as a sick person or a cool person, depending on the context of the communicative scenario. Our emphasis on predicting emerging argument and argument pairing is motivated by a comparison to existing approaches to modeling semantic chaining which use a similar setup for prediction (Grewal & Xu, 2021; Habibi et al., 2020; Xu et al., 2016; Yu & Xu, 2021), and it can be taken as an initial step toward characterizing the general processes of word meaning extension.

Secondly, recent work has shown that chaining models based on semantic proximity between novel and existing meaning (also known as “associative chaining”) often fail to predict metaphorical or other non-literal word meaning extensions (e.g., “to arrive at school” → “to arrive at conclusion”) (Yu, 2023). Future work should explore how infinite mixture models may be integrated with other types of chaining mechanisms to account for mechanisms such as analogy (Fugikawa et al., 2023), which is also discussed in structural mapping (Gentner, 1983) and conceptual metaphor mapping (Lakoff & Johnson, 2008).

Thirdly, our study focused on modeling the cognitive mechanisms that may give rise to novel word choices in lexical evolution, but it does not account for other mechanisms or factors that can also shape the changing landscapes of the lexicon. For example, novel word meanings might emerge due to growing needs for communicating socio-cultural changes or technological innovations, and the chaining models we presented here do not take into account these factors. Related to this issue, we evaluated our framework against historical corpus data that might reflect conventionalized or sustained changes, but this approach is potentially limited in explaining spontaneous or less common changes that do not appear in text corpora. Understanding rare, unconventional patterns of word meaning extension may require a diachronic analysis of linguistic communities, and a characterization of the elimination and propagation of linguistic innovations through the lens of sociolinguistics.

Finally, the current framework of infinite mixture chaining considers word meaning extension at a population level by predicting historically emerging word usages presumably shared among a group of language speakers. We believe that our framework has the potential to explain word meaning extension from individuals. Prior studies have shown that computational models of prototype-based and exemplar-based chaining can explain novel word uses by (individual) children (Pinto & Xu, 2021; Xu & Xu, 2021), and future work can investigate whether infinite mixture chaining might offer an individual-level account of word meaning extension in light of cognitive efficiency.

We have developed a general probabilistic framework for reconstructing emerging word meanings through time and explored a broad class of chaining models that trade off representational accuracy with memory complexity. Our work provides a unified account for the different forms of chaining grounded in that rational models of human and machine learning, and it also opens the avenue for exploring the cognitive efficiency of word meaning acquisition and representation in the mind.

ACKNOWLEDGMENTS

We thank Suzanne Stevenson, the editor, and the anonymous reviewers for their feedback on our work.

FUNDING INFORMATION

This work was supported by an NSERC Discovery Grant RGPIN-2018-05872, a SSHRC Insight Grant 435190272, and an Ontario ERA Award #ER19-15-050 to YX.

AUTHOR CONTRIBUTIONS

LY and YX conceptualized and designed the study, and developed the models. LY acquired data, implemented the models and performed the analyses. LY and YX interpreted the results and wrote the paper. YX acquired funding.

DATA AVAILABILITY STATEMENT

All datasets we used in this study are publicly available and can be downloaded via the following links: Google Syntactic n-gram: https://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html; SemEval 2020 Task 1: https://competitions.codalab.org/competitions/20948; Historical data of Chinese numeral classifiers: https://github.com/AmirAhmadHabibi/ChainingClassifiers.

REFERENCES

Alatrash, R., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2020). CCOHA: Clean corpus of historical American English. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 6958–6966). European Language Resources Association.

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*(5), 789–834. https://doi.org/10.1080/03640210801929287, PubMed: 21635354

Allan, K., & Robinson, J. A. (2012). *Current methods in historical semantics*. De Gruyter Mouton. https://doi.org/10.1515/9783110252903

Allen, K., Shelhamer, E., Shin, H., & Tenenbaum, J. (2019). Infinite mixture prototypes for few-shot learning. In *International conference on machine learning* (pp. 232–241). PMLR.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press. https://doi.org/10.4324/9780203771730

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233. https://doi.org/10.1006/jmps.1995.1021

Blank, A. (1999). Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In A. Blank & P. Koch (Eds.), *Historical semantics and cognition* (pp. 61–90). De Gruyter Mouton. https://doi.org/10.1515/9783110804195.61

Bréal, M. (1897). *Essai de sémantique: Science des significations*. Hachette.

Brochhagen, T., Boleda, G., Gualdoni, E., & Xu, Y. (2023). From language development to language evolution: A unified view of human lexical creativity. *Science*, *381*(6656), 431–436. https://doi.org/10.1126/science.ade7981, PubMed: 37499016

Dąbrowska, E. (2004). Rules or schemas? Evidence from Polish. *Language and Cognitive Processes*, *19*(2), 225–271. https://doi.org/10.1080/01690960344000170

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*(1), e12922. https://doi.org/10.1111/cogs.12922, PubMed: 33432630

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588. https://doi.org/10.1080/01621459.1995.10476550

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209–230. https://doi.org/10.1214/aos/1176342360

Fugikawa, O., Hayman, O., Liu, R., Yu, L., Brochhagen, T., & Xu, Y. (2023). A computational analysis of crosslinguistic regularity in semantic change. *Frontiers in Communication*, *8*, 1136338. https://doi.org/10.3389/fcomm.2023.1136338

Geeraerts, D. (1997). *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford University Press. https://doi.org/10.1093/oso/9780198236528.001.0001

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170. https://doi.org/10.1016/S0364-0213(83)80009-3

Goldberg, Y., & Orwant, J. (2013). A dataset of syntactic-ngrams over time from a very large corpus of English books. In M. Diab, T. Baldwin, & M. Baroni (Eds.), *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity* (pp. 241–247). Association for Computational Linguistics.

Görür, D., & Rasmussen, C. E. (2010). Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, *25*(4), 653–664. https://doi.org/10.1007/s11390-010-9355-8

Grewal, K., & Xu, Y. (2021). Chaining algorithms and historical adjective extension. In N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu, & S. Hengchen (Eds.), *Computational approaches to semantic change* (pp. 189–218). Language Science Press. https://doi.org/10.5281/zenodo.5040312

Griffiths, T. L., Canini, K., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In D. J. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 323–328). Cognitive Science Society.

Habibi, A. A., Kemp, C., & Xu, Y. (2020). Chaining and the growth of linguistic categories. *Cognition*, *202*, 104323. https://doi.org/10.1016/j.cognition.2020.104323, PubMed: 32480166

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489–1501). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1141

Hartigan, J. A., & Wong, M. A. (1979). A *k*-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. https://doi.org/10.2307/2346830

Hilpert, M. (2007). Keeping an eye on the data: Metonymies and their patterns. In A. Stefanowitsch & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 123–151). De Gruyter Mouton. https://doi.org/10.1515/9783110199895.123

Hilpert, M. (2012). Diachronic collostructional analysis: How to use it and how to deal with confounding factors. In K. Allan & J. A. Robinson (Eds.), *Current methods in historical semantics* (pp. 133–160). De Gruyter Mouton. https://doi.org/10.1515/9783110252903.133

Jespersen, O. (1959). *Language: Its nature, development and origin*. Allen & Unwin.

Kay, C., Roberts, J., Samuels, M., Wotherspoon, I., & Alexander, M. (2015). *The historical thesaurus of English, version 4.2*. University of Glasgow.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128. https://doi.org/10.1146/annurev-linguistics-011817-045406

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*(2), 259–282. https://doi.org/10.1006/jmla.2001.2779

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 267–301). Cambridge University Press. https://doi.org/10.1017/CBO9780511816772.013

Kulis, B., & Jordan, M. I. (2012). Revisiting k-means: New algorithms via Bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 1131–1138).

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226471013.001.0001

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago Press.

Li, J., & Joanisse, M. F. (2021). Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science*, *45*(4), e12955. https://doi.org/10.1111/cogs.12955, PubMed: 33873247

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49–70. https://doi.org/10.3758/BF03211715, PubMed: 8433906

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262. https://doi.org/10.1006/jmla.1998.2593

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Nosofsky, R. M. (1986). 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, PubMed: 2937873

Perek, F. (2018). Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, *14*(1), 65–97. https://doi.org/10.1515/cllt-2016-0014

Pinto, R. F., Jr., & Xu, Y. (2021). A computational theory of child overextension. *Cognition*, *206*, 104472. https://doi.org/10.1016/j.cognition.2020.104472, PubMed: 33091729

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, *115*(10), 2323–2328. https://doi.org/10.1073/pnas.1714730115, PubMed: 29463738

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

Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., & Davis, M. H. (2012). Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition*, *40*(7), 1095–1108. https://doi.org/10.3758/s13421-012-0209-1, PubMed: 22614728

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

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*(2), 178–210. https://doi.org/10.1006/jmps.2001.1379

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1–23). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.1

Sloman, S. A., Malt, B. C., & Fridman, A. (2001). Categorization versus similarity: The case of container names. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 73–86). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198506287.003.0005

Sun, Z., & Xu, Y. (2022). Tracing semantic variation in slang. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 1299–1313). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.84

Sun, Z., Zemel, R., & Xu, Y. (2021). A computational framework for slang generation. *Transactions of the Association for Computational Linguistics*, *9*, 462–478. https://doi.org/10.1162/tacl_a_00378

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv*. https://doi.org/10.48550/arXiv.physics/0004057

Traugott, E. C., & Dasher, R. B. (2001). *Regularity in semantic change*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486500

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, *4*(3), 273–290. https://doi.org/10.1515/cogl.1993.4.3.273

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 2277–2282). Lawrence Erlbaum.

Xu, A., & Xu, Y. (2021). Chaining and the formation of spatial semantic categories in childhood. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 700–706). Cognitive Science Society.

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition*, *201*, 104280. https://doi.org/10.1016/j.cognition.2020.104280, PubMed: 32442799

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*(8), 2081–2094. https://doi.org/10.1111/cogs.12312, PubMed: 26456158

Yu, L. (2023). Systematic word meta-sense extension. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 10953–10966). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.675

Yu, L., & Xu, Y. (2021). Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 920–931). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.71

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942. https://doi.org/10.1073/pnas.1800521115, PubMed: 30021851

APPENDIX: RESULTS OF NOUN ARGUMENT PREDICTION PROBLEMS

Additionally to the task of head prediction described in the main text, we also assess the efficiency of different chaining models on the task of predicting novel noun arguments. Figure A1 summarizes the results for noun argument prediction of English head verbs in Case Study 1. We observe that similar to the head prediction problems, among all the models examined, the infinite mixture and exemplar models yielded near-equivalent predictive accuracy and are superior than the alternatives. The prototype and 1nn models perform better than the two baselines, but they are much worse than the top 2 models. Similar trends are also observed in the noun argument prediction problems of English adjectives (Figure A3) and Chinese numeral classifiers (Figure A5).

Figure A2 shows the efficiency trade-off result of noun argument prediction for English verbs in Case Study 1. Again, similar to the trade-off results in head prediction problem, we found that 1) by sweeping λ from 0 toward ∞ (in this case $\lambda \geq 0.5$ suffices), the predictive

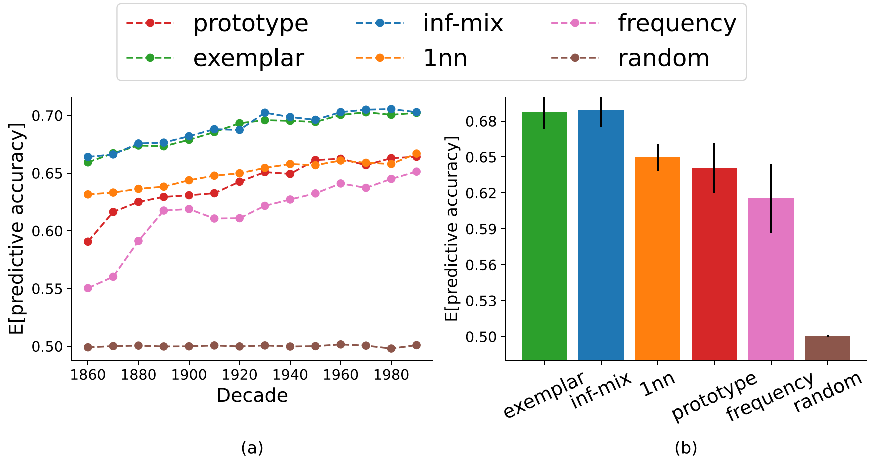


Figure A1. Model accuracy of argument noun prediction for English verbs through time (left panel) and in aggregate (right panel). The infinite mixture model has $\lambda = 0.24$. Error bars represent the standard deviations of accuracy across decades.

accuracy drops only slightly from the exemplar model to the infinite mixture model ($\lambda = 0.24$) but substantially to the prototype model—this finding confirms with our previous analysis, that the infinite mixture model predicts on par with the exemplar model; and 2) the marginal gain on accuracy of the exemplar model comes at a high cost in memory complexity: compared to the infinite-mixture model, it requires over 2^5-fold more storage of cluster centroids to achieve a gain of <0.01 in predictive accuracy. We observed a similar trade-off curve for the argument noun prediction problems of English adjectives (Figure A4) and Chinese numeral classifiers (Figure A6) as well. Overall, the infinite mixture model achieves a better balance between precision and memory.

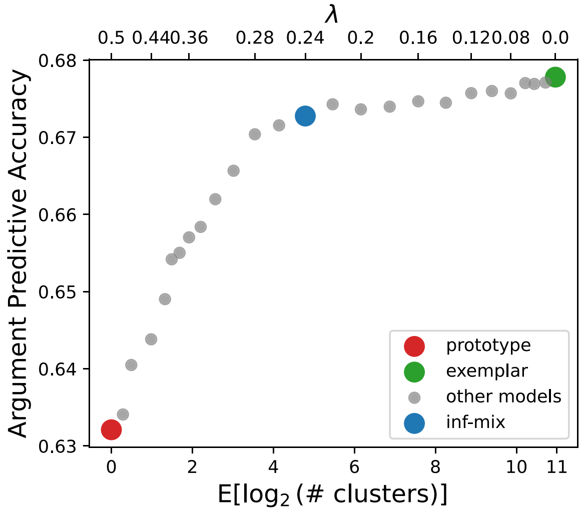


Figure A2. Accuracy of noun argument prediction problem and memory complexity (mean number of clusters per word used in prediction) on English historical verb-noun composition dataset. The gray dots show the spectrum of infinite mixture models under different λ values. The left end (red dot) of the x-axis corresponds to the prototype model with $\lambda \geq 0.5$. The right end (green dot) of the x-axis corresponds to the exemplar model with $\lambda = 0$. The blue dot corresponds to the infinite mixture model with inferred optimal λ value.

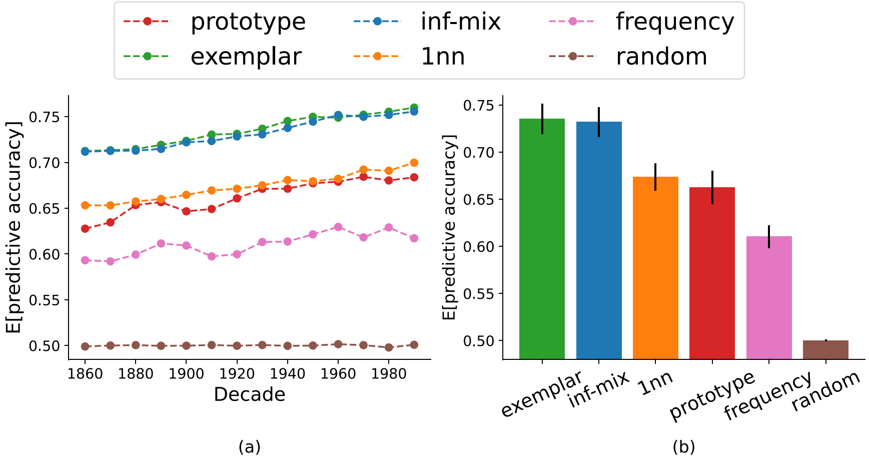


Figure A3. Model accuracy of argument noun prediction for English adjectives through time (left panel) and in aggregate (right panel). The infinite mixture model has $\lambda = 0.14$. Error bars represent the standard deviations of accuracy across decades.

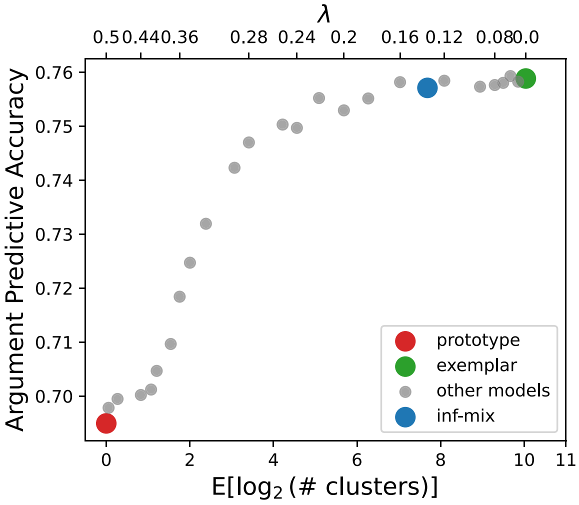


Figure A4. Accuracy of noun argument prediction problem and memory complexity (mean number of clusters per word used in prediction) on English historical adjective-noun composition dataset.

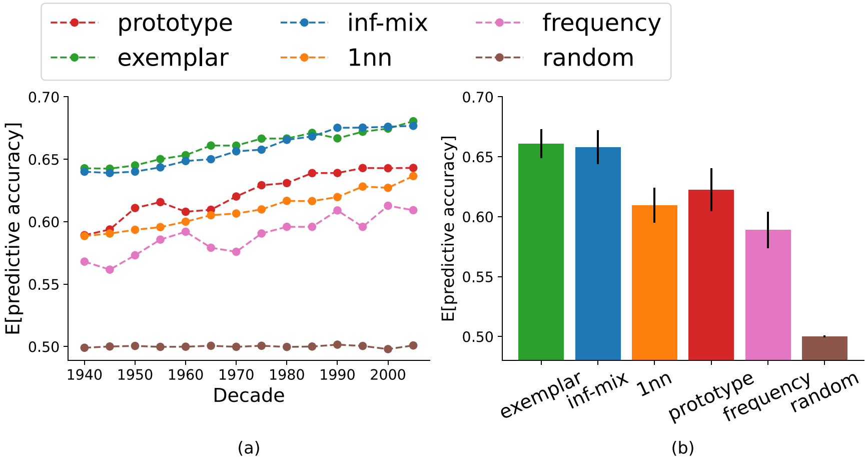


Figure A5. Model accuracy in argument noun prediction for Chinese numeral classifiers through time (left panel) and in aggregate (right panel). The infinite mixture model has λ = 3.0. Error bars represent the standard deviations of accuracy across decades.

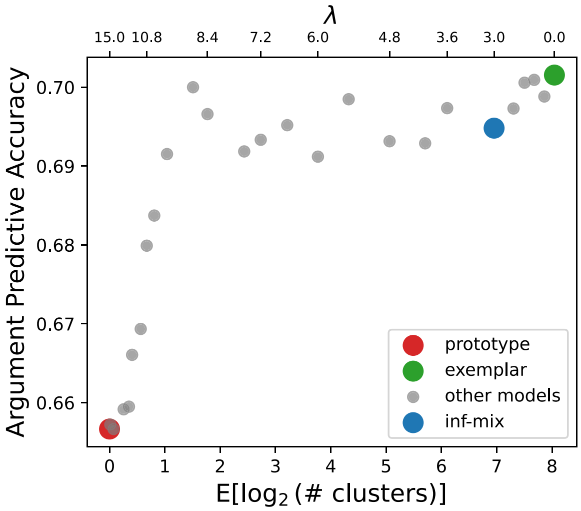


Figure A6. Accuracy of noun argument problem and memory complexity (mean number of clusters per word used in prediction) on Chinese historical noun-classifier composition dataset.