Infinite mixture chaining: Efficient temporal construction of word meaning

Lei Yu (jadeleiyu@cs.toronto.edu)
Department of Computer Science, University of Toronto

Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science, Cognitive Science Program, University of Toronto

Abstract

Word meanings extend over time due to a functional need for maintaining communicative expressivity within a compact lexicon. Previous scholars have suggested that word meanings extend via a process of chaining, whereby novel items link to existing ones close in semantic space. Recent work has formalized this idea using computational models grounded typically in the exemplar and prototype theories of categorization that are either memory-intensive or simplistic in representation. We propose an alternative account of chaining that optimizes cognitive efficiency by trading off representational accuracy with memory complexity. We operationalize this efficient chaining as an infinite mixture model and show how it constructs the internal representations of word meaning adaptively through time while predicting the historical development of English verb meanings with precision and limited resources.

Keywords: the lexicon; historical semantics; word meaning extension; infinite mixture chaining; cognitive efficiency

Introduction

Words often take on new meanings. For example, the noun *face* in English referred to "body part" earlier but later extended to "facial expression" and "front surface of an object" (from *Historical Thesaurus of English*) . Similarly, the verb *store* progressively took on emerging items like *food*, *electricity*, and *password* as its noun arguments over the past centuries (see Figure 1). C. S. Lewis vividly pictured word meaning extension as "a tree throwing out new branches", a historical process he referred to as "ramification" (Lewis, 1990). Wittgenstein also described the ramification of this process as "family resemblance": how a word embraces a polysemous set of meanings forming "a complicated network of similarities overlapping and crisscrossing" (Wittgenstein, 1953). More generally, linguists have suggested that language change results from a functional need for maximizing communicative expressivity under minimum effort (Jespersen, 1959; Blank, 2013). Indeed, recent work offered empirical support to this view suggesting that word meaning extension is a dominant strategy for maintaining expressivity of the lexicon toward emerging meanings while keeping it compact (Ramiro, Srinivasan, Malt, & Xu, 2018). What are the cognitive mechanisms that support the flexible construction of novel word meanings over time? Here we investigate this question in a formal framework that explores the processes of word meaning extension through the lens of cognitive efficiency.

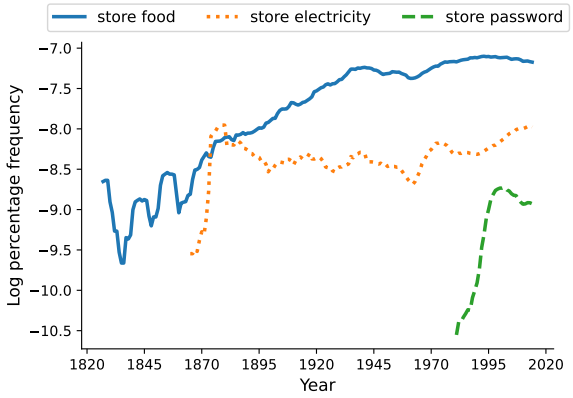
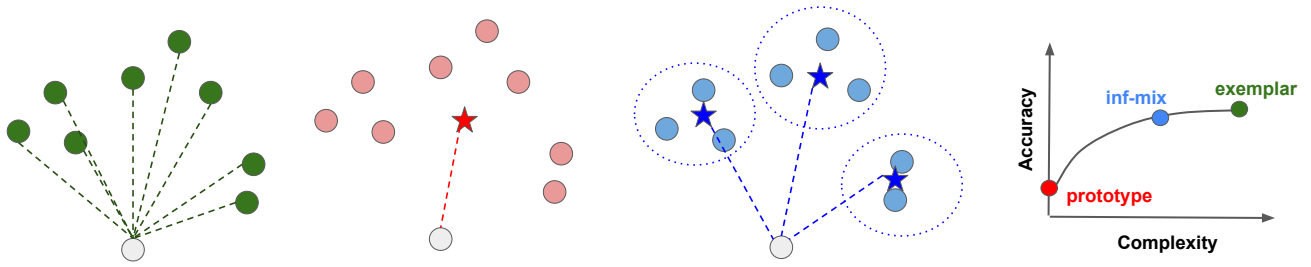


Figure 1: Usage frequencies of the phrases *store food*, *store electricity*, and *store password* in the past 200 years of English. Data from Google Syntactic-Ngrams historical corpus.

A prominent view on the process of word meaning extension originates from scholars in cognitive linguistics and psychology. By this view, word meanings extend via a process of chaining, whereby new items tend to link to existing meanings of a word when they are proximal in semantic space, resulting in chain-like structures over time (Lakoff, 1987; Malt, Sloman, Gennari, Shi, & Wang, 1999; Hilpert, 2008). Recent work has extended this view and developed formal models of chaining to explain the historical extension of container names (Sloman, Malt, & Fridman, 2001; Xu, Regier, & Malt, 2016), numeral classifiers (Habibi, Kemp, & Xu, 2020), adjectives (Grewal & Xu, 2021), verb frames (Yu & Xu, 2021), informal word usages (Sun, Zemel, & Xu, 2021), and word senses in general (Ramiro et al., 2018). All of these studies have focused on two main types of chaining mechanism, grounded typically either in the tradition of a prototype model which assumes that each lexical category is represented by a central prototype (Reed, 1972; Rosch, 1975; Lakoff, 1987), or in terms of an exemplar-based model which assumes that each category is represented by its set of exemplars stored in memory (Nosofsky, 1986; Ashby & Alfonso-Reese, 1995). Which of these models best describes chaining has received mixed views, although it has been suggested that the exemplar-based approach tends to predict historical data better than the prototype model (Habibi et al., 2020). How-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*



(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).

ever, a common assumption in this line of research is that the prototype and exemplar-based models are adequate to capture the chaining processes in word meaning extension. Here we challenge this assumption under the view that mechanisms of chaining should be assessed under the consideration of cognitive efficiency (Jespersen, 1959)—an important aspect that has not been explored comprehensively in the existing work.

We define cognitive efficiency in word meaning extension as processes that operate under two competing constraints that trade off against each other: *representational accuracy* and *memory complexity*. The prototype and exemplar-based approaches to chaining described fall under two extremes in this tradeoff. At one extreme, the exemplar model offers a highly accurate representation of the mental state of a (lexical) category by storing the past exemplars, and it can 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 necessary expense of a high memory cost. 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 necessary expense of a simplistic if not impoverished representation, hence why it tends to suffer from inferior predictability in historical word meaning extension (Habibi et al., 2020). This exemplar-prototype dichotomy can thus be understood as an intrinsic tradeoff in cognitive efficiency: An accurate model often demands a high memory load, while a minimum-effort model tends to be poor in representational precision. The open issue is whether there are alternative accounts of chaining that near-optimally achieve cognitive efficiency.

Research in rational human and machine learning has suggested a third possibility that near-optimally trades off the two competing dimensions of efficiency. Under this view, a lexical category can be modelled as an infinite mixture of clusters of exemplars (see Figures 2c); critically this clustering scheme can be flexibly adjusted to capture the internal structure of a category as it assimilates new items (Anderson, 1990; Rosseel, 2002; Vanpaemel, Storms, & Ons, 2005; Griffiths, Canini, Sanborn, & Navarro, 2007). In our case, an infinite mixture approach to chaining can potentially help represent polysemy (Klein & Murphy, 2001; Rodd et al., 2012; Tuggy, 1993) and complex structures of word meaning over time beyond the exemplar and prototype models which either represent word meanings as a set of independent exemplars or a prototype. Similar views have been proposed in statistical machine learning often in the tradition of Dirichlet process (DP) mixture (Ferguson, 1973; Escobar & West, 1995; Allen, Shelhamer, Shin, & Tenenbaum, 2019) which instantiates a tradeoff between information loss (in model reconstruction of data) and complexity (in terms of the number of clusters inferred by model) (Kulis & Jordan, 2012).

Here we propose a general theoretical framework of chaining that explicitly takes into account how different models behave on the accuracy-complexity tradeoff plane (see Figures 2d). Our framework relates to a growing body of research suggesting natural language is structured to support efficient communication that trades off informativeness and complexity (Kirby, Tamariz, Cornish, & Smith, 2015; Kemp, Xu, & Regier, 2018; Zaslavsky, Kemp, Regier, & Tishby, 2018; Gibson et al., 2019). However, our study also differs from this line of work by grounding the temporal mechanisms of chaining in the notion of cognitive efficiency. Our framework, dubbed *infinite mixture chaining* (abbreviated as *inf-mix*), offers a new way of constructing word meanings dynamically as they emerge through time. It does so by automatically forming semantically related clusters represented by their centroids for joint memory and representation efficiency. We show that our framework subsumes both prototype and exemplar models under the variation of a single

tradeoff parameter, and the infinite mixture model of chaining predicts historical data equally well as the exemplar model while requiring a lower memory complexity. For the scope of this study, we focus on predicting historical verb meaning extension as verbs acquire new noun arguments through time, illustrated in Figure 1.

Theoretical framework

We formulate word meaning extension under a probabilistic framework by focusing on verbs as an exemplary case, but we expect this framework to generalize similarly to other word classes. In the following, we first consider meaning extension as a temporal prediction problem under the two constraints of cognitive efficiency. We then show how several classes of chaining models can be derived from this framework and describe the diachronic semantic space in which these models are operationalized.

Problem formulation under efficiency constraints

We define word meaning extension as a temporal inference problem. Given a word and its current meaning at time t, we wish to infer which novel items will likely emerge into that word's referential range in the near future. In the case of verb meaning extension, we cast this problem as probabilistic inference over novel verb-noun compositions over time. Specifically, given a verb v such as *store*, we ask which noun arguments can be paired with that verb to form previously unattested compositions that extend its meaning space, e.g., *store*:"food"→"electricity"→ "password". Since a verb can take noun arguments under different syntactic roles (e.g., direct object, or dobj vs. subject), we also constrain syntactic relation r in predicting verb-noun compositions. Formally, we consider a verb-relation pair (v,r) (e.g., *store in dobj*) as a category denoted by $\mathcal{S}_{v,r}$, and the temporal inference problem is equivalent to predicting the probability of any query noun n_q to emerge in that category at a future time. We focus on predicting n_q's that have not yet appeared as noun arguments for a given verb, i.e., novel verb-noun compositions. For instance, the category "*store in dobj*" may have been attested to pair with the noun *food* up to time t, but predicted to extend toward new nouns such as *information* later.

Given a list of previously unattested query noun arguments $n_q \in Q^t_{v,r}$ at time t, our framework infers which nouns will be appropriate arguments for verb v under syntactic relation r at time $t+\Delta$ where Δ is an increment in time:

$$p(n_q|v,r)^{t+\Delta} = p(n_q|\mathcal{S}^t_{v,r}) \propto \text{sim}(n_q, \mathcal{S}^t_{v,r}) \quad (1)$$

Here $\text{sim}(n_q, \mathcal{S}^t_{v,r})$ is a yet-to-be-specified function (i.e., different ways of chaining) that measures the semantic similarity between the query noun and current meaning of the verb-relation category $\mathcal{S}_{v,r}$ at time t. To compute this similarity, we quantify the semantic proximity between n_q and the existing set of noun arguments of $\mathcal{S}_{v,r}$ (i.e., category exemplars). We refer to this set of nouns as the support set (denoted by $n_s \in \mathcal{S}_{v,r}$). 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 $\mathcal{M}_{v,r}$.

$$\text{sim}(n_q, \mathcal{S}^t_{v,r}) = \text{sim}(n_q, \mathcal{M}^t_{v,r}) = \text{sim}(n_q, \{\mu^t_{v,r,k}\}_{k=1}^{K^t_{v,r}}) \quad (2)$$

Here $\mathcal{M}^t_{v,r} = \{\mu^t_{v,r,1}, \mu^t_{v,r,2}, ...\}$ is a set of $K^t_{v,r}$ cluster centroids for support set $\mathcal{S}^t_{v,r}$. In the next section, we show that exemplar chaining is equivalent to the case where each support noun (or exemplar) is in 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 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 verb-relation category:

$$\text{sim}(n_q, \mathcal{M}^t_{v,r}) = \frac{1}{K^t_{v,r}} \sum_{k=1}^{K^t_{v,r}} \exp(-\|\phi(n_q)^t - \mu^t_{v,r,k}\|^2) \quad (3)$$

We allow the number of cluster centroids to flexibly vary over time (as a verb encounters new nouns), which is inferred and updated based on the internal semantic structure of a verb-relation category instantiated in terms of its support nouns. In particular, the semantic clusters inferred within a category are expected to optimize the following tradeoff between two constraints of efficiency, following work on infinite mixtures from machine learning (Kulis & Jordan, 2012):

$$\mathcal{M}^t_{v,r} = \text{argmin}_{\mathcal{M}} \sum_{k}^{K^t_{v,r}} \sum_{n_s \in S^t_{v,r}} \|\phi(n_s)^t - \mu^t_k\|^2 + \lambda K^t_{v,r} \quad (4)$$

The first term on the right of Equation 4 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). 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

Decade	Verb-relation pair		Support noun	Query noun
	Predicate verb	Syntactic relation		
1900	drive	direct object	horse, wheel, cart	car, van
1950	work	prepositional object via *as*	mechanic, carpenter, scientist	astronaut, programmer
1980	store	prepositional object via *in*	fridge, container, box	supercomputer

Table 1: Sample entries from Google Syntactic-Ngrams including verb-relation pairs, support and query nouns, and timestamps.

communication, which assumes that word meanings are organized under the tradeoff between reconstruction accuracy and complexity (Tishby, Pereira, & Bialek, 2000; Zaslavsky et al., 2018). However, a crucial distinction is that here our emphasis is in model inference of newly emerging meanings for individual words rather than (synchronic or retro) construction of a semantic system.

Classes of chaining model

The efficiency formulation in Equation 4 helps derive several classes of chaining model from the literature and anew, and we show that our framework subsumes these classes.

Exemplar-based models. In the case where the tradeoff parameter $\lambda \to 0$, 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:

$$p(n_q|v,r)^{t+\Delta} \propto \frac{1}{|S^t_{v,r}|} \sum_{n_s \in S^t_{v,r}} \exp(-\|\phi(n_q)^t - \phi(n_s)^t\|^2) \quad (5)$$

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:

$$p(n_q|v,r)^{t+\Delta} \propto \text{argmax}_{n_s \in S^t_{v,r}} \exp(-\|\phi(n_q)^t - \phi(n_s)^t\|^2) \quad (6)$$

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]

$$p(n_q|v,r)^{t+\Delta} \propto \exp(-\|\phi(n_q)^t - \mu^t_{v,r}\|^2) \quad (7)$$

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

[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.

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 $|S^t_{v,r}|$ 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 $\mathcal{M}^t_{v,r}$ 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 $p(n_q|v,r)$. 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 mixture centroids μ_k become latent variables and need to be inferred via posterior sampling, which requires storing all support noun arguments and is computationally prohibitive. Our framework bypasses these issues of DPGMM and is more computationally tractable.

Diachronic semantic space

The chaining models described need to be operationalized in a time-varying semantic space so that information about future verb-noun usages 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, Sutskever, Chen, Corrado, & Dean, 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 pretrained diachronic embeddings (Hamilton, Leskovec, & Jurafsky, 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 linguistic data (Yu & Xu, 2021), which can provide alternative semantic representations.

Data

To evaluate our framework, we collected a large dataset of historical verb-noun compositions derived from the Google Syntactic-Ngrams (GSN) English corpus (Lin et al., 2012) from 1850 to 2000. Table 1 shows sample entries of data

which we will make publicly available.[2] Specifically, we collected verb-noun-relation triples $(n,v,r)^t$ that co-occur in the ENGALL subcorpus of GSN over the 150 years. 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 should have at least $\theta_v = 20,000$ counts in GSN. To support feasible computations, we consider the top-20 most common syntactic relations in GSN, such as direct object, direct subject, and relations concerning prepositional objects. 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 pairs over 14 decades, where each verb-relation category has at least 1 novel query noun and 10 existing support nouns.

Results

We evaluated different classes of chaining models under variation of the tradeoff parameter λ on predicting emerging verb-noun compositions for the historical period 1850 to 2000. At every decade, for each verb-relation pair (v,r) with a query noun n_q, we randomly sample 100 alternative noun arguments from the vocabulary of top-10,000 nouns in GSN that never appeared with v under relation r in the corpus, 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. This procedure allows us to assess the degree to which each class of chaining model can successfully predict novel verb-noun pairings incrementally through time, and how they fair in the accuracy-complexity tradeoff.

For infinite mixture models with $0 < \lambda < \infty$, we implemented the DP-means clustering algorithm introduced in Kulis & 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.

[2] https://github.com/jadeleiyu/inf-mix-chaining

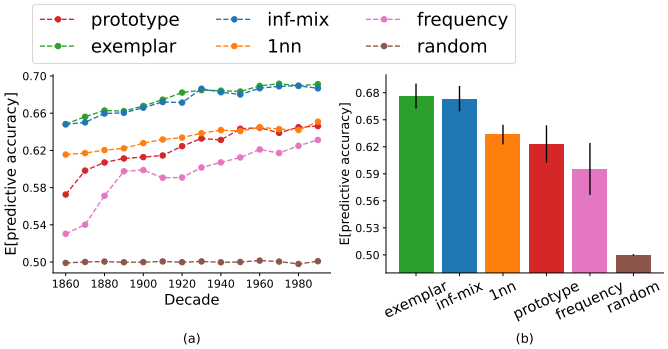


Figure 3: Model accuracy in predicting emerging verb-noun compositions through time (left panel) and in aggregate (right panel). The infinite mixture model has $\lambda = 0.24$. Error bars represent the standard deviations of accuracy across decades.

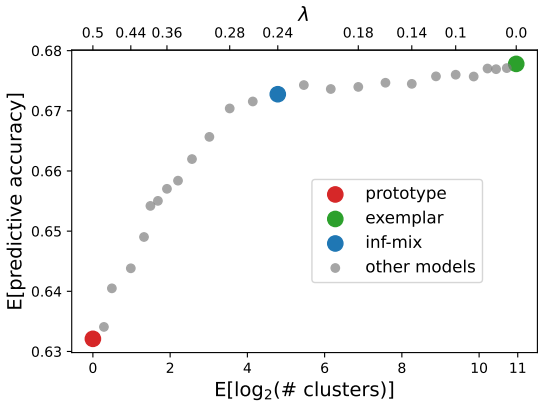


Figure 4: Predictive accuracy (averaged over 150 years) and memory complexity (mean number of clusters per word used in prediction). The gray dots show the spectrum of infinite mixture models under different λ values. The left end (red dot) of the x-axis corresponds to the prototype model with $\lambda \geq 0.5$. 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.

Temporal prediction of verb-noun compositions. We found that when $\lambda = 0.24$, the inf-mix model yields most well-defined clustering overall measured by the standard Silhouette score for unsupervised clustering.[3] 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 3 summarizes the results. We observe

[3] 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.

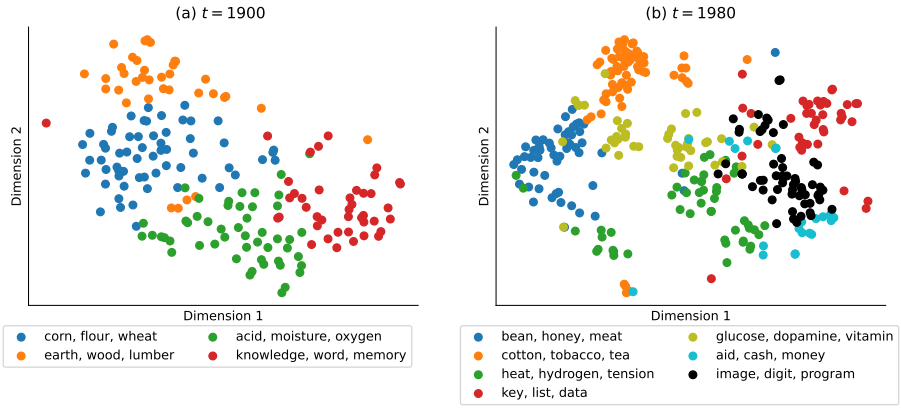


Figure 5: Low-dimensional visualizations of historical meaning extension for the verb frame *store in (noun)* from 1900s (left) to 1980s (right) via t-SNE projection. The dots correspond to word embeddings of noun arguments grouped in clusters inferred by the infinite mixture chaining model. Legends show 3 representative nouns closest to their cluster centroids for each cluster.

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. 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 dominantly performing model as reported in recent work of chaining (Habibi et al., 2020; Yu & Xu, 2021). We next assess whether how the infinite mixture model fairs against the exemplar model in memory complexity and efficiency.

The accuracy-complexity tradeoff. 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 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 4 shows the results 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 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. Overall, the infinite mixture model achieves a better balance between precision and memory.

Interpretation. We interpret the semantic clusters learned by the infinite mixture model using verb category *store in dobj* as an example. Figure 5 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 5. 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).

Conclusion

We have presented a unified framework of semantic chaining, examined through a large dataset of historical verb-noun compositions. We show how existing accounts of chaining fall under this coherent framework governed by a tradeoff parameter that modulates the competing constraints of representational accuracy and memory complexity. The infinite mixture model predicts emergent verb-noun compositions equally well as the exemplar model but at a lower memory cost. Our work suggests that word meanings are constructed over time in cognitively efficient ways and builds the first link among theories of chaining, efficiency of natural language, and rational models of human and machine learning.

Acknowledgments

We would like to thank members of the Cognitive Lexicon Laboratory at University of Toronto for their constructive comments on this work. This work was supported by a NSERC Discovery Grant RGPIN-2018-05872, a SSHRC Insight Grant #435190272, and an Ontario ERA Award to YX.

References

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).

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, *39*(2), 216–233.

Blank, A. (2013). 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.

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

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 209–230.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Goldberg, Y., & Orwant, J. (2013, June). A dataset of syntactic-ngrams over time from a very large corpus of English books. In *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). Atlanta, Georgia, USA: 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.

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). Berlin: Language Science Press.

Griffiths, T., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical dirichlet process.

Habibi, A. A., Kemp, C., & Xu, Y. (2020). Chaining and the growth of linguistic categories. *Cognition*, *202*, 104323.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1489–1501).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100–108.

Hilpert, M. (2008). Keeping an eye on the data: Metonymies and their patterns. In *Corpus-based approaches to metaphor and metonymy* (pp. 123–151). De Gruyter Mouton.

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

Kay, C., Roberts, J., Samuels, M., Wotherspoon, I., & Alexander, M. (2015). *The Historical Thesaurus of English, version 4.2*. Glasgow, UK: University of Glasgow.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*(2), 259–282.

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.

Lewis, C. S. (1990). *Studies in words*. Cambridge University Press.

Lin, Y., Michel, J.-B., Lieberman, E. A., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the acl 2012 system demonstrations* (pp. 169–174).

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.

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.

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.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

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.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*(2), 178–210.

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). New York: Oxford University Press.

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

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, *4*(3).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In *Proceedings of the annual conference of the cognitive science society* (Vol. 27, pp. 2277–2282).

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Basil Blackwell.

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.

Yu, L., & Xu, Y. (2021). Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 920–931).

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. Retrieved from `http://www.pnas.org/content/115/31/7937` doi: 10.1073/pnas.1800521115