Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining

Lei Yu[1], **Yang Xu**[1,2,3]
[1] Department of Computer Science, University of Toronto, Toronto, Canada
[2] Cognitive Science Program, University of Toronto, Toronto, Canada
[3] Vector Institute for Artificial Intelligence, Toronto, Canada
{jadeleiyu,yangxu}@cs.toronto.edu

Abstract

Natural language relies on a finite lexicon to express an unbounded set of emerging ideas. One result of this tension is the formation of new compositions, such that existing linguistic units can be combined with emerging items into novel expressions. We develop a framework that exploits the cognitive mechanisms of chaining and multimodal knowledge to predict emergent compositional expressions through time. We present the syntactic frame extension model (SFEM) that draws on the theory of chaining and knowledge from "percept", "concept", and "language" to infer how verbs extend their frames to form new compositions with existing and novel nouns. We evaluate SFEM rigorously on the 1) modalities of knowledge and 2) categorization models of chaining, in a syntactically parsed English corpus over the past 150 years. We show that multimodal SFEM predicts newly emerged verb syntax and arguments substantially better than competing models using purely linguistic or unimodal knowledge. We find support for an exemplar view of chaining as opposed to a prototype view and reveal how the joint approach of multimodal chaining may be fundamental to the creation of literal and figurative language uses including metaphor and metonymy.

1 Introduction

Language users often construct novel compositions through time, such that existing linguistic units can be combined with emerging items to form novel expressions. Consider the expression *swipe your phone*, which presumably came about after the emergence of touchscreen-enabled smartphones. Here the use of the verb *swipe* was extended to express one's experience with the emerging item "smartphone". These incremental extensions are fundamental to adapting a finite lexicon toward emerging communicative needs. We explore the nature of cognitive mechanisms and knowledge in the temporal formation of previously unattested verb-argument compositions, and how this compositionality may be understood in principled terms.

Compositionality is at the heart of linguistic creativity yet a notoriously challenging topic in computational linguistics and natural language processing (e.g., Vecchi et al. 2017; Cordeiro et al. 2016; Blacoe and Lapata 2012; Mitchell and Lapata 2010; Baroni and Zamparelli 2010). For instance, modern views on the state-of-the-art neural models of language have suggested that they show some degree of linguistic generalization but are impoverished in systematic compositionality (see Baroni (2020) for review). Existing work has also explored the efficacy of neural models in modeling diachronic semantics (e.g., Hamilton et al., 2016; Rosenfeld and Erk, 2018; Hu et al., 2019; Giulianelli et al., 2020). However, to our knowledge, no attempt has been made to examine principles in the formation of novel verb-noun compositions through time.

We formulate the problem as an inferential process which we call *syntactic frame extension*. We define syntactic frame as a joint distribution over a verb predicate, its noun arguments, and their syntactic relations, and we focus on tackling two related predictive problems: 1) given a novel or existing noun, infer what verbs and syntactic relations that have not predicated the noun might emerge to describe it over time (e.g., *to drive a car* vs. *to fly a car*), and 2) given a verb predicate and a syntactic relation, infer what nouns can be plausibly introduced as its novel arguments in the future (e.g., *drive a car* vs. *drive a computer*).

Figure 1 offers a preview of our framework by visualizing the process of assigning novel verb frames to describe two query nouns over time. In the first case, the model incorporated with perceptual and conceptual knowledge successfully predicts the verb *drive* to be a better predicate than *fly* for describing the novel item *car* that just emerged at the time of prediction where linguistic usages are not yet observed (i.e., emergent verb compo-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 920–931
November 7–11, 2021.

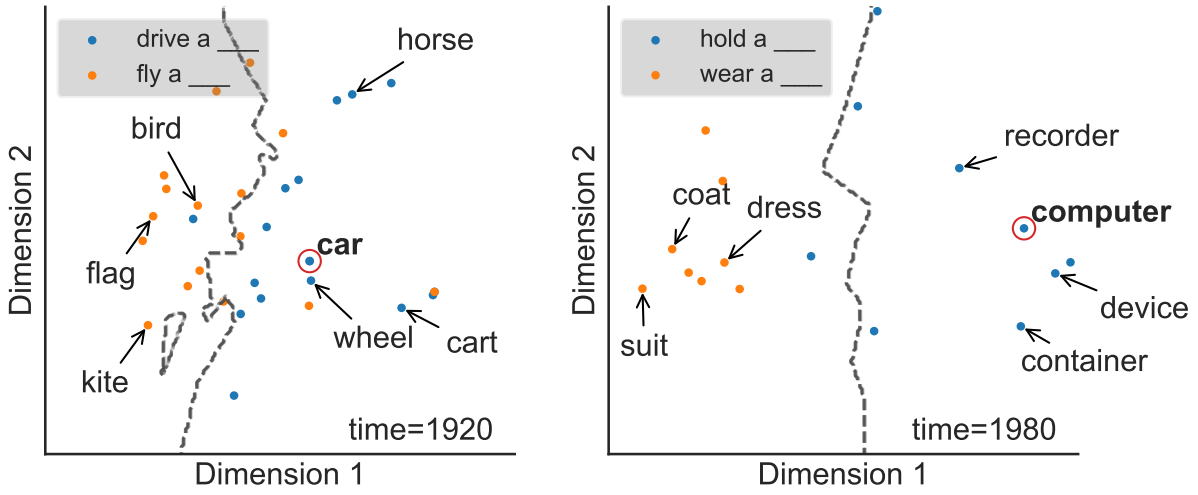

Figure 1: Preview of the proposed approach of syntactic frame extension. Given a query noun (circled dot) at time t, the framework draws on a combination of multimodal knowledge and cognitive mechanisms of chaining to predict novel linguistic expressions for those items. The left panel shows via PCA projection how a newly emerged noun (i.e., *car* in 1920s) is assigned appropriate verb frames by comparing the learned multimodal representation of the query nouns with support nouns (non-circled dots) that have been predicated by the frames. The right panel shows a similar verb construction for a noun that already existed at the time of prediction (i.e., *computer* in 1980s).

sition with a novel noun concept). In the second case, the model predicts that *hold* is a better predicate than *wear* for describing the noun *computer*, which already existed at the time of prediction (i.e., emergent verb composition with an existing noun).

Our approach connects two strands of research that were rarely in contact: cognitive linguistic theories of chaining and computational representations of multimodal semantics. Work in the cognitive linguistics tradition has suggested that frame extension is not arbitrary and involves the comparison between a new item to existing items that are relevant to the frame (Fillmore, 1986). Similar proposals lead to the theory of chaining postulating that linguistic categories grow by linking novel referents to existing ones of a word due to proximity in semantic space (Lakoff, 1987; Malt et al., 1999; Xu et al., 2016; Ramiro et al., 2018; Habibi et al., 2020; Grewal and Xu, 2020). However, such a theory has neither been formalized nor evaluated to predicting verb frame extensions through time. Separately, computational work in multimodal semantics has suggested how word meanings warrant a richer representation beyond purely linguistic knowledge (e.g., Bruni et al. 2012; Gella et al. 2016, 2017). However, multimodal semantic representations have neither been examined in the diachronics of compositionality nor in light of the cognitive theories of chaining. We show that a unified framework that incorporates the cognitive mechanisms of chaining through deep models of categorization and multimodal semantic representations predicts the temporal emergence of novel noun-verb compositions.

2 Related work

Our work synthesizes the interdisciplinary areas of cognitive linguistics, diachronic semantics, meaning representation, and deep learning.

2.1 Cognitive mechanisms of chaining

The problem of syntactic frame extension concerns the cognitive theory of chaining (Lakoff, 1987; Malt et al., 1999). It has been proposed that the historical growth of linguistic categories depends on a process of chaining, whereby novel items link to existing referents of a word that are close in semantic space, resulting in chain-like structures. Recent studies have formulated chaining as models of categorization from classic work in cognitive science. Specifically, it has been shown that chaining may be formalized as an exemplar-based mechanism of categorization emphasizing semantic neighborhood profile (Nosofsky, 1986), which contrasts with a prototype-based mechanism that emphasizes category centrality (Reed, 1972; Lakoff, 1987; Rosch, 1975). This computational approach to chaining has been applied to explain word meaning growth in numeral classifiers (Habibi et al., 2020) and adjectives (Grewal and Xu, 2020). Unlike these pre-

vious studies, we consider the open issue whether cognitive mechanisms of chaining might be generalized to verb frame extension which draws on rich sources of knowledge. It remains critically undetermined how "shallow models" such as the exemplar model can function or integrate with deep neural models (Mahowald et al., 2020; McClelland, 2020), and how it might fair with the alternative mechanism of prototype-based chaining in the context of verb frame extension. We address both of these theoretical issues in a framework that explores these alternative mechanisms of chaining in light of probabilistic deep categorization models.

2.2 Diachronic semantics in NLP

The recent surge of interest in NLP on diachronic semantics has developed Bayesian models of semantic change (e.g., Frermann and Lapata, 2016), diachronic word embeddings (e.g., Hamilton et al., 2016), and deep contextualized language models (e.g., Rosenfeld and Erk, 2018; Hu et al., 2019; Giulianelli et al., 2020). A common assumption in these studies is that linguistic usages (from historical corpora) are sufficient to capture diachronic word meanings. However, previous work has suggested that text-derived distributed representations tend to miss important aspects of word meaning, including perceptual features (Andrews et al., 2009; Baroni and Lenci, 2008; Baroni et al., 2010) and relational information (Necşulescu et al., 2015). It has also been shown that both relational and perceptual knowledge are essential to construct creative or figurative language use such as metaphor (Gibbs Jr. et al., 2004; Gentner and Bowdle, 2008) and metonymy (Radden and Kövecses, 1999). Our work examines the function of multimodal semantic representations in capturing diachronic verb-noun compositions, and the extent to which such representations can be integrated with the cognitive mechanisms of chaining.

2.3 Multimodal representation of meaning

Computational research has shown the effectiveness of grounding language learning and distributional semantic models in multimodal knowledge beyond linguistic knowledge (Lazaridou et al., 2015; Hermann et al., 2017). For instance, Kiros et al. (2014) proposed a pipeline that combines image-text embedding models with LSTM neural language models. Bruni et al. (2014) identifies discrete "visual words" in images, so that the distributional representation of a word can be extended to encompass its co-occurrence with the visual words of images it is associated with. Gella et al. (2017) also showed how visual and multimodal information help to disambiguate verb meanings. Our framework extends these studies by incorporating the dimension of time into exploring how multimodal knowledge predicts novel language use.

2.4 Memory-augmented deep learning

Our framework also builds upon recent work on memory-augmented deep learning (Vinyals et al., 2016; Snell et al., 2017). In particular, it has been shown that category representations enriched by deep neural networks can effectively generalize to few-shot predictions with sparse input, hence yielding human-like abilities in classifying visual and textual data (Pahde et al., 2020; Singh et al., 2020; Holla et al., 2020). In our work, we consider the scenario of constructing novel compositions as they emerge over time, where sparse linguistic information is available. We therefore extend the existing line of research to investigate how representations learned from naturalistic stimuli (e.g., images) and structured knowledge (e.g., knowledge graphs) can reliably model the emergence of flexible language use that expresses new knowledge and experience.

3 Computational framework

We present the syntactic frame extension model (SFEM), which is composed of two components. First, SFEM specifies a frame as a joint probabilistic distribution over a verb, its noun arguments, and their syntactic relations and supports temporal prediction of verb syntax and arguments via deep probabilistic models of categorization. Second, SFEM draws on multimodal knowledge by incorporating perceptual, conceptual, and linguistic cues into flexible inference for extended verb frames over time. Figure 2 illustrates our framework.

3.1 Chaining as probabilistic categorization

We denote a predicate verb as v (e.g., *drive*) and a syntactic relation as r (e.g., direct object of a verb), and consider a finite set of verb-syntactic frame elements $f = (v, r) \in \mathcal{F}$. We define the set of nouns that appeared as arguments for a verb (under historically attested syntactic relations) up to time t as support nouns, denoted by $n_s \in S(f)^{(t)}$ (e.g., *horse* appeared as a support noun—the direct object—for the verb *drive* prior to 1880s). Given a query noun n^* (e.g., *car* upon its emergence in

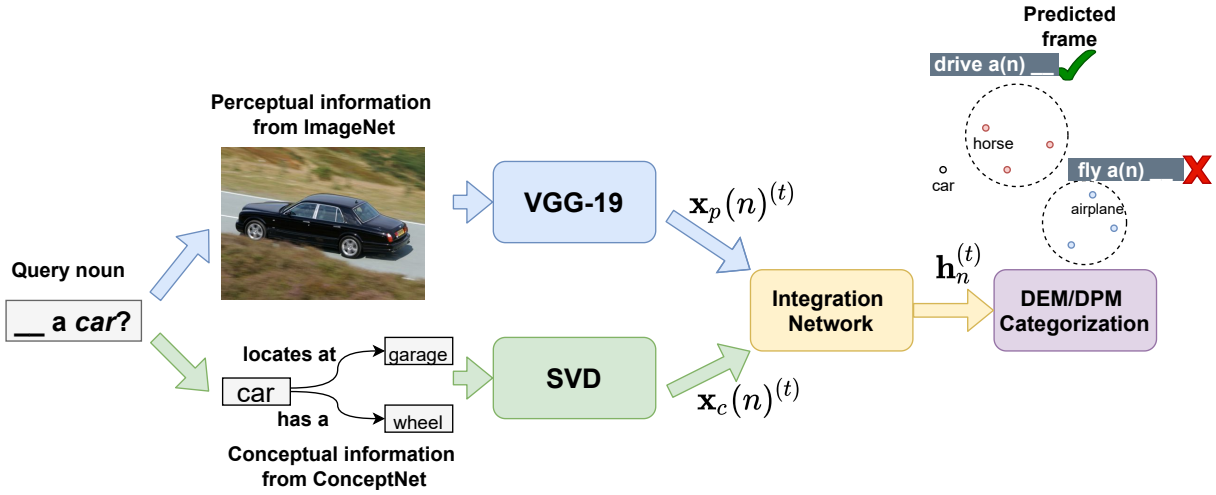

Figure 2: Illustration of the syntactic frame extension model for the emerging query *car*. The model integrates information from visual perception and conceptual knowledge about cars to form a multimodal embedding ($h_n^{(t)}$), which supports temporal prediction of appropriate verb-syntax usages via deep categorization models of chaining.

1880s) that has never been an argument of v under relation r, we define syntactic frame extension as probabilistic inference in two related problems:

1. **Verb syntax prediction.** Here we predict which verb-syntactic frames f are appropriate to describe the query noun n^*, operationalized as $p(f|n^*)$ yet to be specified.

2. **Noun argument prediction.** Here we predict which nouns n^* are plausible novel arguments for a given verb-syntax frame f, operationalized as $p(n^*|f)$ yet to be specified.

We solve these inference problems by modeling the joint probability $p(n^*, f)$ for a query noun and a candidate verb-syntactic frame incrementally through time as follows:

$$p(n^*, f)^{(t)} = p(n^*|f)^{(t)} p(f)^{(t)} \quad (1)$$

$$= p(n^*|S(f)^{(t)}) p(f)^{(t)} \quad (2)$$

Here we construct verb meaning based on its existing support nouns $S(f)^{(t)}$ at current time t. We infer the most probable verb-syntax usages for describing the query noun (Problem 1) as follows:

$$p(f|n^*) = \frac{p(n^*, f)^{(t)}}{\sum_{f \in \mathcal{F}} p(n^*, f)^{(t)}} \quad (3)$$

$$= \frac{p(n^*|S(f)^{(t)}) p(f)^{(t)}}{\sum_{f' \in \mathcal{F}} p(n^*|S(f')^{(t)}) p(f')^{(t)}} \quad (4)$$

In the learning phase, we train our model incrementally at each time period t by minimizing the log joint probability $p(n^*, f)^{(t)}$ in Equation 1 for every frame f and each of its query noun $n^* \in Q(f)^{(t)}$:

$$J = -\sum_{f \in \mathcal{F}^{(t)}} \sum_{n^* \in Q(f)^{(t)}} \log p(n^*, f)^{(t)} \quad (5)$$

For each noun n, we consider a time-dependent hidden representation $\mathbf{h}_n^{(t)} \in \mathbb{R}^M$ derived from different sources of knowledge (specified in Section 3.2). For the prior probability $p(f)^{(t)}$, we consider a frequency-based approach that computes the proportion for the number of unique noun arguments that a (verb) frame has been paired with and attested in a historical corpus:

$$p(f)^{(t)} = \frac{|S(f)^{(t)}|}{\sum_{f' \in \mathcal{F}} |S(f')^{(t)}|} \quad (6)$$

We formalize $p(n^*|f)$ (Problem 2), namely $p(n^*|N(f))^{(t)}$, by two classes of deep categorization models motivated by the literature on chaining, categorization, and memory-augmented learning.

Deep prototype model (SFEM-DPM). SFEM-DPM draws inspirations from prototypical network for few-shot learning (Snell et al., 2017) and is grounded in the prototype theory of categorization in cognitive psychology (Rosch, 1975). The model computes a set of hidden representations for every support noun $n_s \in S(f)^{(t)}$, and takes the expected vector as a *prototype* $\mathbf{c}_f^{(t)}$ to represent f at time t:

$$\mathbf{c}_f^{(t)} = \frac{1}{|S(f)^{(t)}|} \sum_{n_s \in S(f)^{(t)}} \mathbf{h}_{n_s}^{(t)} \quad (7)$$

The likelihood of extending n^* to $S(f)^{(t)}$ is then defined as a softmax distribution over l_2 distances $d(\cdot,\cdot)$ to the embedded prototype:

$$p(n^*|S(f)^{(t)}) = \frac{\exp\left(-d(\mathbf{h}_{n^*}^{(t)}, \mathbf{c}_f^{(t)})\right)}{\sum_{f'} \exp(-d(\mathbf{h}_{n^*}^{(t)}, \mathbf{c}_f^{(t)}))} \quad (8)$$

Deep exemplar model (SFEM-DEM). In contrast to the prototype model, SFEM-DEM resembles the memory-augmented matching network in deep learning (Vinyals et al., 2016), and formalizes the exemplar theory of categorization (Nosofsky, 1986) and chaining-based category growth (Habibi et al., 2020). Unlike DPM, this model depends on the l_2 distances between n^* and every support noun:

$$p(n^*|S(f)^{(t)}) = \frac{\sum_{n_s \in S(f)^{(t)}} \exp\left(-d(\mathbf{h}_{n^*}^{(t)}, \mathbf{h}_{n_s}^{(t)})\right)}{\sum_{f'} \sum_{n'_s \in S(f')^{(t)}} \exp\left(-d(\mathbf{h}_{n^*}, \mathbf{h}_{n'_s}^{(t)})\right)} \quad (9)$$

3.2 Multimodal knowledge integration

In addition to the probabilistic formulation, SFEM draws on structured knowledge including perceptual, conceptual, and linguistic cues to construct multimodal semantic representations $\mathbf{h}_n^{(t)}$ introduced in Section 3.1.

Perceptual knowledge. We capture perceptual knowledge from image representations in the large, taxonomically organized ImageNet database (Deng et al., 2009). For each noun n, we randomly sample a collection of 64 images from the union of all ImageNet synsets that contains n, and encode the images through the VGG-19 convolutional neural network (Simonyan and Zisserman, 2015) by extracting the output vector from the last fully connected layer after all convolutions (see similar procedures also in Pinto Jr. and Xu, 2021). We then average the encoded images to a mean vector $\mathbf{x}_p(n) \in \mathbb{R}^{1000}$ as the perceptual representation of n.

Conceptual knowledge. To capture conceptual knowledge beyond perceptual information (e.g., attributes and functions), we extract information from the ConceptNet knowledge graph (Speer et al., 2017), which connects concepts in a network structure via different types of relations as edges. This graph reflects commonsense knowledge of a concept (noun) such as its functional role (e.g., a car IS_USED_FOR *transportation*), taxonomic information (e.g., a car IS_A *vehicle*), or attributes (e.g., a car HAS_A *wheel*). Since the concepts and their relations may change over time, we prepare a diachronic slice of the ConceptNet graph at each time t by removing all words with frequency up to t in a reference historical text corpus (see Section 4 for details) under a threshold k_c which we set to be 10. We then compute embeddings for the remaining concepts following methods recommended in the original study by Speer et al. (2017). In particular, we perform singular value decomposition (SVD) on the positive pointwise mutual information matrix $\mathbf{M}_G^{(t)}$ of the ConceptNet $G^{(t)}$ truncated at time t, and combine the top 300 dimensions (with largest singular values) of the term and context matrix symmetrically into a concept embedding matrix. Each row of the resulting row matrix of SVD will therefore serves as the conceptual embedding $\mathbf{x}_c(n)^{(t)} \in \mathbb{R}^{300}$ for its corresponding noun.

Linguistic knowledge. For linguistic knowledge, we take the HistWords diachronic word embeddings $\mathbf{x}_l^{(t)} \in \mathbb{R}^{300}$ pre-trained on the Google N-Grams English corpora to represent linguistic meaning of each noun at decade t (Hamilton et al., 2016).

Knowledge integration. To construct a unified representation that incorporates knowledge from different modalities, we take the mean of the unimodal representations described into a joint vector $\mathbf{x}_n \in \mathbb{R}^{300}$, and then apply an integration function $g : \mathbb{R}^{300} \to \mathbb{R}^M$ parameterized by a feedforward neural network to get the multimodal word representation $h_n^{(t)}$.[1] Our framework allows flexible combinations of the three modalities introduced, e.g., a full model would utilizes all three types of knowledge, while a linguistic-only baseline will directly take HistWords embeddings $\mathbf{x}_l^{(t)}$ as inputs of the integration network.

4 Historical noun-verb compositions

To evaluate our framework, we collected a large dataset of historical noun-verb compositions derived from the Google Syntactic N-grams (GSN) English corpus (Lin et al., 2012) from 1850 to 2000. 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) a noun n

[1] For ImageNet embeddings, we apply a linear transformation to project each $\mathbf{x}_p^{(t)}$ into $\mathbb{R}^{300}$ so that all unimodal representations are 300-d vectors before taking the means.

Decade	Verb syntactic frame		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 Ngram including verb syntactic frames, support and query nouns.

should have at least $\theta_p = 64$ image representations in ImageNet, $\theta_c = 10$ edges in the contemporary ConceptNet network, and $\theta_n = 15,000$ counts (with POS tag as nouns) in GSN over the 150-year period; 2) a verb v should have at least $\theta_v = 15,000$ counts in GSN. To facilitate feasible model learning, we consider the top-20 most common syntactic relations in GSN, including 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 query nouns n^* for a given verb frame f if their number of co-occurrences with f up to time t falls below a threshold θ_q, while the number of co-occurrences with f up to time $t + \Delta$ is above θ_q (i.e., an emergent use that conventionalizes). We define support nouns as those that co-occurred with f 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 10,349 verb-syntactic frames over 15 decades, where each frame class has at least 1 novel query noun and 4 existing support nouns. Table 1 shows sample entries of data which we make publicly available.[2]

5 Evaluation and results

We first describe the details of SFEM implementation and diachronic evaluation. We then provide an in-depth analysis on the multimodal knowledge and chaining mechanisms in verb frame extension.

5.1 Details of model implementation

We implemented the integration network $g(\cdot)$ of SFEM as a three-layer feedforward neural network with an output dimension $M = 100$, and keep parameters and embeddings in other modules fixed during learning.[3] At each decade, we randomly sample 70% of the query nouns with their associated verb-syntactic pairs as training data, and take the remaining examples for model testing such that there is no overlap in the query nouns between training and testing. We trained models on the negative log-likelihood loss defined in Equation 5 at each decade. To examine how multimodal knowledge contributes to temporal prediction of novel language use, we trained 5 DEM and 5 DPM models using information from different modalities.

5.2 Evaluation against historical data

We test our models on both verb syntax and noun argument predictive tasks with the goals of assessing 1) the contributions of multimodal knowledge, and 2) the two alternative mechanisms of chaining. We also consider baseline models that do not implement chaining-based mechanisms: a frequency baseline that predicts by count in GSN up to time t, and a random guesser. We evaluate model performance via standard receiver operating characteristics (ROC) curves that reveal cumulative precision of models in their top m predictions. We compute the standard area-under-curve (AUC) statistics for the ROC curves to get the mean precision over all values of m from 1 to the candidate set size. Figure 3 summarizes the results over the 150 years. We observe that 1) all multimodal models perform better than their uni-/bi-modal counterparts, and 2) the exemplar-based model performs dominantly better than the prototype-based counterpart, and both outperform the baseline models without chaining. In particular, a tri-modal deep exemplar model that incorporates knowledge from all three modalities achieves the best overall performance. These results provide strong support that verb frame extension depends on multimodal knowledge and an exemplar-based chaining.

To further assess how the models perform in predicting emerging verb extension toward both novel and existing nouns, we report separate mean AUC scores for these predictive cases where query nouns are either completely novel (i.e., zero token frequencies) or established. (i.e., above-zero frequencies) at the time of prediction. Table 2 summarizes

[2]Data and code are deposited here: https://github.com/jadeleiyu/frame_extension

[3]See Appendix A for additional implementation details.

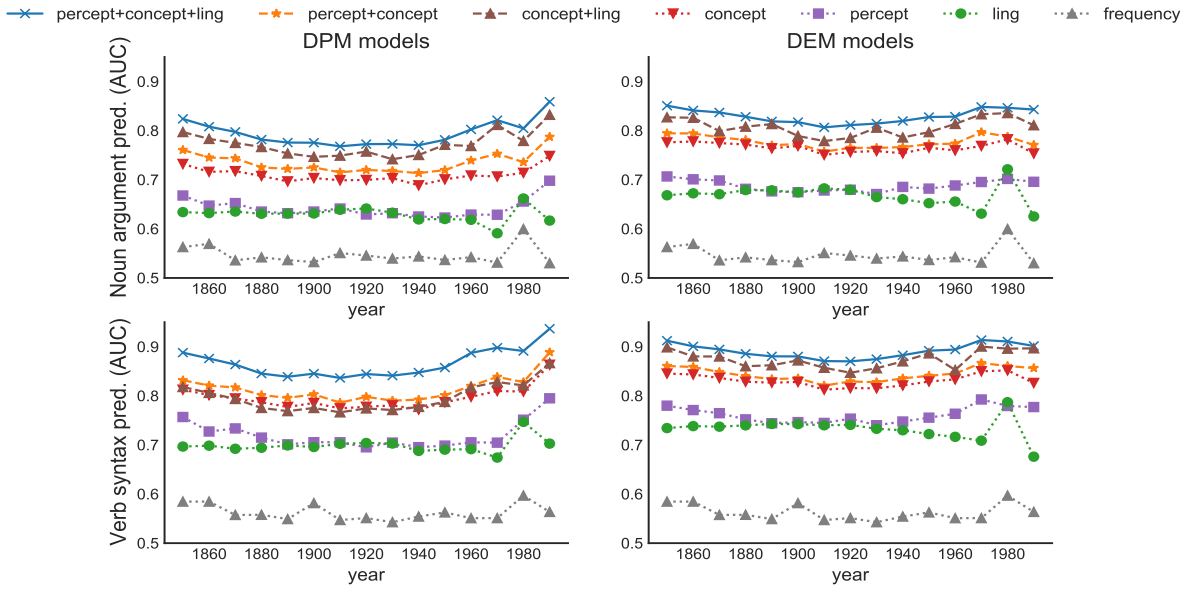

Figure 3: Area-under-curves of SFEM and baseline models from 1850s to 1990s. Top row: AUCs of predicting extended syntax frames for query nouns. Bottom row: AUCs of predicting extended nouns for query verb frames.

these results and shows that model performances are similar under four predictive cases. For prediction with novel query nouns, it is not surprising that linguistic-only models fail due to the unavailability of linguistic mentions. However, for prediction with established query nouns, the superiority of multimodal SFEMs is still prominent suggesting that our framework captures general principles in verb frame extension (and not just for predicting verb extension toward novel nouns).

Table 3 compares sample verb syntax predictions made by the full and linguistic-only DEM models that cover a diverse range of concepts including inventions (e.g., airplane), discoveries (e.g., microorganism), and occupations (e.g. astronaut). We observe that the full model typically constructs reasonable predicate verbs that reflect salient features of the query noun (e.g., *cars* are vehicles that are *drive*-able). In contrast, the linguistic-only model often predicts verbs that are either overly generic (e.g., *purchase* a telephone) or nonsensical.

5.3 Model analysis and interpretation

We provide further analyses and interpret why both multimodality and chaining mechanisms are fundamental to predicting emergent verb compositions.

5.3.1 The function of multimodal knowledge

To understand the function of multimodal knowledge, we compute, for each modality, the top-4 verb compositions that were most degraded in joint probability $p(f, n)$ after ablating a knowledge modality from the full tri-modal SFEM (see Table 4). A drop in $p(f, n)$ indicates reliance on multimodality in prediction. We found that linguistic knowledge helps the model identify some general properties that are absent in the other cues (e.g., a *monitor* is *buy*-able). Importantly, for the two extra-linguistic knowledge modalities, we observe that visual-perceptual knowledge helps predict many imaged-based metaphors, including "the airplane rolls" (i.e., based on common shapes of airplanes) and "the tree stands" (based on verticality of trees). On the other hand, conceptual knowledge predicts cases of logical metonymy (e.g., "work for the newspaper") and conceptual metaphor (e.g., "kill the process"). These examples suggest that multimodality serves to ground and embody SFEM with commonsense knowledge that constructs novel verb compositions for not only literal language use, but also non-literal or figurative language use that is extensively discussed in the psycholinguistics literature (Lakoff, 1982; Radden and Kövecses, 1999; Gibbs Jr. et al., 2004).

We also evaluate the contributions of the three modalities in model prediction by comparing the AUC scores from the three uni-modal DEMs. Figure 4 shows the percentage breakdown of examples on which one of the modalities yields the highest score (i.e., contributes most to a reliable prediction). We observe that conceptual cues explain data the best in almost 2/3 of the cases, followed by percep-

Model	AUC – verb syntax prediction			AUC – noun argument prediction		
	novel items	existing items	combined	novel items	existing items	combined
DPM (linguistics)	0.642	0.690	0.681	0.641	0.653	0.650
DPM (perceptual)	0.632	0.666	0.657	0.650	0.624	0.629
DPM (conceptual)	0.772	0.722	0.733	0.727	0.705	0.711
DPM (perceptual+conceptual)	0.809	0.754	0.767	0.725	0.719	0.721
DPM (perceptual+linguistics)	0.645	0.669	0.661	0.655	0.669	0.665
DPM (conceptual+linguistics)	0.753	0.774	0.766	0.776	0.768	0.770
DPM (perceptual+conceptual+linguistics)	0.848	0.810	0.815	0.799	0.786	0.788
DEM (linguistics)	0.652	0.690	0.686	0.641	0.625	0.632
DEM (perceptual)	0.737	0.674	0.684	0.659	0.650	0.655
DEM (conceptual)	0.854	0.784	0.788	0.736	0.724	0.729
DEM (perceptual+conceptual)	0.858	0.792	0.797	0.750	0.744	0.748
DEM (perceptual+linguistics)	0.712	0.759	0.753	0.698	0.710	0.708
DEM (conceptual+linguistics)	0.902	0.866	0.870	0.837	**0.822**	0.824
DEM (perceptual+conceptual+linguistics)	**0.919**	**0.872**	**0.878**	**0.856**	0.820	**0.827**
Baseline (frequency)	0.573	0.573	0.573	0.536	0.536	0.536
Baseline (random)	0.500	0.500	0.500	0.500	0.500	0.500

Table 2: Mean model AUC scores of verb syntax and noun argument predictions from 1850s to 1990s.

Query noun	Decade	Predicted frames (linguistic-only DEM)	Predicted frames (tri-modal DEM)
telephone	1860	roll-nsubj, load-dobj, play-pobj_prep.on	purchase_dobj, pick-dobj, remain-pobj_prep.on
microorganism	1900	decorate-pobj_prep.with, play-pobj_prep.on, spread-dobj	feed-pobj_prep.on, mix-pobj_prep.with, breed-dobj
airplane	1930	load-dobj, mount-pobj_prep.on, mount-dobj, blow-dobj, roll-nsubj	fly-dobj, approach-nsubj, drive-dobj, stop-nsubj
astronaut	1950	spin-dobj,work-pobj_prep.in, emerge-pobj_prep.from	work-pobj_prep.as, talk-pobj_prep.to, lead-pobj_prep.by
computer	1970	purchase-dobj, fix-dobj, generate-dobj, write-pobj_prep.to	store-pobj_prep.in, move-pobj_prep.into, display-pobj_prep.on, implement-pobj_prep.in

Table 3: Example predictions of novel verb-noun compositions from the full tri-imodal and linguistic-only models.

Ablated modality	Most affected compositions
Language	buy a monitor, find a disk (*), a resident dies, specialized in nutrition (*), point to the window
Percept	an airplane rolls, talk to an entrepreneur (*), the tree stands, the doctor says, topped with nutella (*)
Concept	perform in the film (*), work as a programmer (*), work for the newspaper, expand the market, kill the process

Table 4: Top-4 ground-truth compositions with most prominent drops in joint probability $p(f, n)$ after ablation of one modality of knowledge from SFEM. Phrases marked with '*' include novel query nouns.

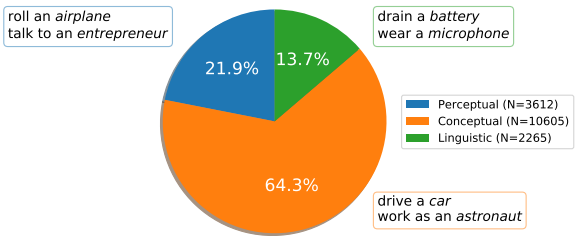

Figure 4: Percentage breakdown of the three modalities in model prediction, with annotated examples.

tual and linguistic cues. These results suggest that while conceptual knowledge plays a dominant role in model prediction, all three modalities contain complementary information in predicting novel language use through time.

5.3.2 General mechanisms of chaining

We next analyze general mechanisms of chaining by focusing on understanding the superiority of exemplar-based chaining in SFEM. Figure 5 illus-

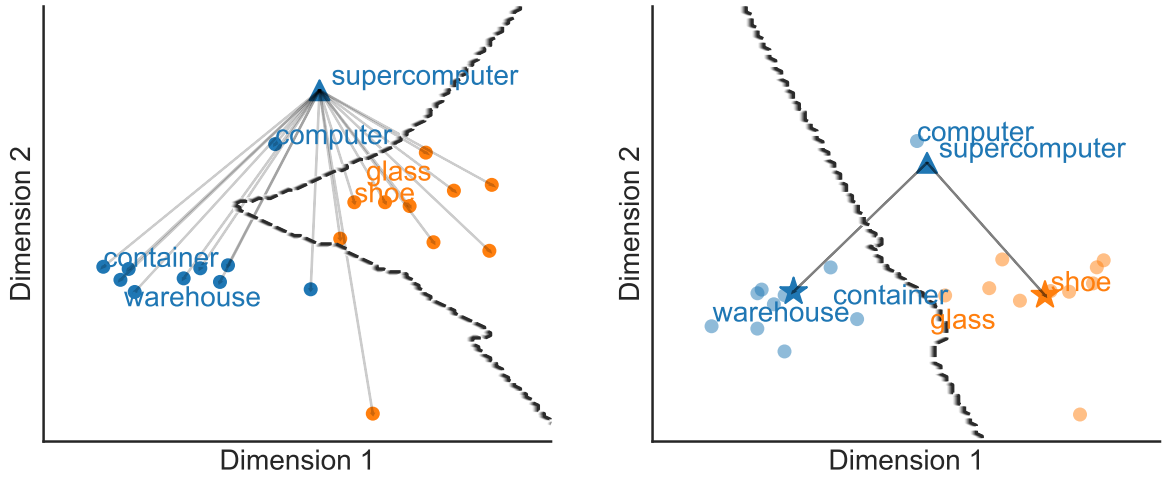

Figure 5: Illustrations of two mechanisms of chaining (left: exemplar; right: prototype) in verb frame prediction for query *supercomputer*. Nouns are PCA-projected in 2D, with categories color-coded and in dashed boundaries.

trates the exemplar-based and prototype-based processes of chaining with the example verb frame prediction for the noun "supercomputer". For simplicity, we only show two competing frames "to store in a ___" and "to wear a ___". In this case, the query noun is semantically distant to most of the prototypical support nouns in both categories, and is slightly closer to the centroid of the "wear" class than to that of the "store" class. The prototype model would then predict the incorrect composition "to wear a supercomputer". In contrast, the exemplar model is more sensitive to the semantic neighborhood profile of the query noun and the aprototypical support noun "computer" of the "store in ___" class, and it therefore correctly predicts that "supercomputer" is more likely to be predicated by "to store in". Our discovery that the exemplar-based chaining accounts for verb composition through time mirrors existing findings on similar mechanisms of chaining in the extensions of numeral classifiers (Habibi et al., 2020) and adjectives (Grewal and Xu, 2020), and together they suggest a general cognitive mechanism may underlie historical linguistic innovation.

6 Conclusion

We have presented a probabilistic framework for characterizing the process of syntactic frame extension in which verbs extend their referential range toward novel and existing nouns over time. Our results suggest that language users rely on extralinguistic knowledge from percept and concept to construct new linguistic compositions via a process of exemplar-based chaining. Our work creates a novel approach to diachronic compositionality and strengthens the link between multimodal semantics and cognitive linguistic theories of categorization.

Acknowledgments

We would like to thank Graeme Hirst, Michael Hahn, and Suzanne Stevenson for their feedback on the manuscript, and members of the Cognitive Lexicon Laboratory at the University of Toronto for helpful suggestions. We also thank the anonymous reviewers for their constructive comments. 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

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463.

Marco Baroni. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307.

Marco Baroni and Alessandro Lenci. 2008. Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55–88.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A distributional semantic model based on property and types. *Cognitive Science*, 34(2):222–254.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1183–1193.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 546–556.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Charles J Fillmore. 1986. Frames and the semantics of understanding. *Quaderni di Semantica*, 6:222–254.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Spandana Gella, Frank Keller, and Mirella Lapata. 2017. Disambiguating visual verbs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):311–322.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–192, San Diego, California. Association for Computational Linguistics.

Dedre Gentner and Brian Bowdle. 2008. Metaphor as structure-mapping. *The Cambridge handbook of metaphor and thought*, 109:128.

Raymond W Gibbs Jr., Paula Lenz Costa Lima, and Edson Francozo. 2004. Metaphor is grounded in embodied experience. *Journal of Pragmatics*, 36(7):1189–1210.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973.

Karan Grewal and Yang Xu. 2020. Chaining and historical adjective extension. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

Amir Ahmad Habibi, Charles Kemp, and Yang Xu. 2020. Chaining and the growth of linguistic categories. *Cognition*, 202:104323.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 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)*, pages 1489–1501.

Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, et al. 2017. Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*.

Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Learning to learn to disambiguate: Meta-learning for few-shot word sense disambiguation. *arXiv preprint arXiv:2004.14355*.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 595–603.

George Lakoff. 1982. Experiential factors in linguistics. *Language, mind, and brain*, pages 142–157.

George Lakoff. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174.

Kyle Mahowald, George Kachergis, and Michael C Frank. 2020. What counts as an exemplar model, anyway? a commentary on ambridge (2020). *First Language*, 40(5-6):608–611.

Barbara C Malt, Steven A Sloman, Silvia Gennari, Meiyi Shi, and Yuan Wang. 1999. Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, 40(2):230–262.

James L McClelland. 2020. Exemplar models are useful and deep neural networks overcome their limitations: A commentary on ambridge (2020). *First Language*, 40(5-6):612–615.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Silvia Necşulescu, Sara Mendes, David Jurgens, Núria Bel, and Roberto Navigli. 2015. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 182–192.

Robert M Nosofsky. 1986. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39.

Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. 2020. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2644–2653.

Renato Ferreira Pinto Jr. and Yang Xu. 2021. A computational theory of child overextension. *Cognition*, 206:104472.

Günter Radden and Zoltán Kövecses. 1999. Towards a theory of metonymy. *Metonymy in language and thought*, 4:17–60.

Christian Ramiro, Mahesh Srinivasan, Barbara C Malt, and Yang Xu. 2018. Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, 115(10):2323–2328.

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

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

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Pulkit Singh, Joshua C Peterson, Ruairidh M Battleday, and Thomas L Griffiths. 2020. End-to-end deep prototype and exemplar models for predicting human behavior. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society (CogSci 2020)*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Eva M Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2017. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science*, 41(1):102–136.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29:3630–3638.

Yang Xu, Terry Regier, and Barbara C Malt. 2016. Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, 40(8):2081–2094.

A Additional details of SFEM implementation

We implemented the integration network $g(\cdot)$ as a three-layer feedforward neural network using PyTorch, where each layer has a dimension of $300, 200$ and 100 respectively. For models that incorporates less than three modalities, we replace the missing embeddings with a zero vector when computing the mean vectors before knowledge integration.

During training, except for network weights in $g(\cdot)$, we keep parameters in every modules (i.e., the VGG-19 encoder and every unimodal embedding) constant, and optimize SFEM by minimzing the negative log-likelihood loss function specified in Equation 5 via stochastic gradient descent (SGD).

Each training batch consists of $B = 64$ syntactic frames with their associated query and support nouns. We train each model for 200 epochs and save the configuration that achieves the highest validation accuracy for our evaluation described in Section 5.