Noun2Verb: Probabilistic Frame Semantics for Word Class Conversion

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

Yang Xu
University of Toronto
Department of Computer Science
Cognitive Science Program
Vector Institute for Artificial Intelligence
yangxu@cs.toronto.edu

Humans can flexibly extend word usages across different grammatical classes, a phenomenon known as word class conversion. Noun-to-verb conversion, or denominal verb (e.g., to Google a cheap flight), is one of the most prevalent forms of word class conversion. However, existing natural language processing systems are impoverished in interpreting and generating novel denominal verb usages. Previous work has suggested that novel denominal verb usages are comprehensible if the listener can compute the intended meaning based on shared knowledge with the speaker. Here we explore a computational formalism for this proposal couched in frame semantics. We present a formal framework, Noun2Verb, that simulates the production and comprehension of novel denominal verb usages by modeling shared knowledge of speaker and listener in semantic frames. We evaluate an incremental set of probabilistic models that learn to interpret and generate novel denominal verb usages via paraphrasing. We show that a model where the speaker and listener cooperatively learn the joint distribution over semantic frame elements better explains the empirical denominal verb usages than state-of-the-art language models, evaluated against data from (1) contemporary English in both adult and child speech, (2) contemporary Mandarin Chinese, and (3) the historical development of English. Our work grounds word class conversion in probabilistic frame semantics and bridges the gap between natural language processing systems and humans in lexical creativity.

1. Introduction

Word class conversion refers to the extended use of a word across different grammatical classes without overt changes in word form. Noun-to-verb conversion, or denominal

Action Editor: Saif Mohammad. Submission received: 12 August 2021; revised version received: 4 March 2022; accepted for publication: 30 March 2022.

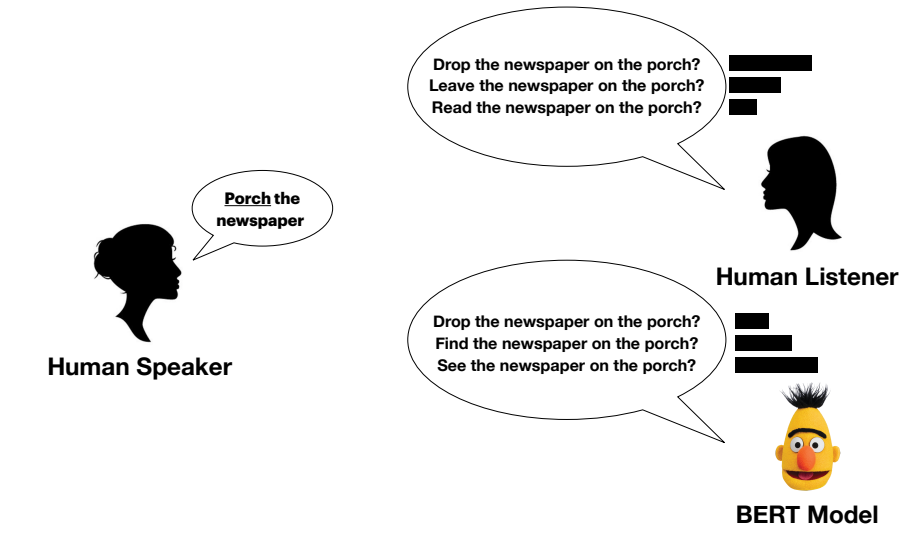
https://doi.org/10.1162/coli_a_00447

verb, is one of the most commonly observed forms of word class conversion. For instance, the expression *to Google a cheap flight* illustrates the innovative verb usage of *Google*, which is conventionally a noun denoting the Web search engine or company. The extended verb use here signifies the action of "searching information online." Although denominal verbs have been studied extensively in linguistics as a phenomenon of lexical semantic innovation in adults and children and across different languages (e.g., Clark and Clark 1979; Clark 1982; Vogel and Comrie 2011; Jespersen 2013), they have been largely underexplored in the existing literature of computational linguistics; and their flexible nature presents key challenges to natural language understanding and generation of innovative word usages. We present a formal computational account of noun-to-verb conversion couched in frame semantics. We show how our probabilistic framework yields sensible interpretation and generation of novel denominal verb usages that go beyond state-of-the-art language models in natural language processing.

Previous work has offered extensive empirical investigations into when noun-to-verb conversion occurs from the viewpoints of syntax (Hale and Keyser 1999), semantics (Dirven 1999), and pragmatics (Clark and Clark 1979). In particular, Clark and Clark (1979) present one of the most comprehensive studies on this topic and describe "the innovative denominal verb convention" as a communicative scenario where the listener can readily comprehend the meaning of a novel denominal verb usage based on Grice's cooperative principles (Grice 1975). They suggest that the successful comprehension of a novel or previously unobserved denominal verb usage relies on the fact that the speaker denotes the kind of state, event, or process that they believe the listener can readily and uniquely compute on the basis of their mutual knowledge. They illustrate this idea with the classic example *the boy porched the newspaper* (see also Figure 1a). Upon hearing this utterance that features the novel denominal use of *porch*, the listener is expected to identify the scenario of a boy delivering the newspaper onto a porch, based on the shared world knowledge about the entities invoked by the utterance: the boy, the porch, and newspaper delivery systems.

In contrast to human language users, existing natural language processing systems often fail to interpret (or generate) flexible denominal utterances in sensible ways. Figure 1a illustrates this problem in two established natural language processing systems. In Figure 1a, a state-of-the-art BERT language model assigned higher probabilities to two inappropriate paraphrases for the query phrase *to porch the newspaper* over the more reasonable paraphrase *to drop the newspaper on the porch*. In Figure 1b, the Google Translate system also failed to back-translate the same query denominal utterance from Mandarin Chinese to English. Specifically, this system misinterpreted the denominal verb "to porch" with the translation "to confuse" in Mandarin Chinese, which led to the erroneous back-translation into English. These failed cases demonstrate the challenges toward natural language processing systems in interpreting flexible denominal verb usages, and they suggest that a principled computational methodology for supporting automated interpretation and generation of novel denominal verb usages may be warranted.

Work from cognitive linguistics, particularly frame semantics, provides a starting point for tackling this problem from the view of structured meaning representation. Specifically, frame semantics theory asserts that humans understand word meaning by accessing a coherent mental structure of encyclopedic knowledge, or *semantic frames*, that store a complex series of events, entities, and scenarios along with a group of participants (Fillmore 1968). Similar conceptual structures have also been discussed by researchers in artificial intelligence, cognitive psychology, and linguistics, under the different terminologies of schema (Minsky 1974; Rumelhart 1975), script (Schank 1972),



(a)

(b)

Figure 1
Illustrations of the problem of noun-to-verb conversion, or denominal verb, in human language users and natural language processing systems. (a) Given a novel denominal usage of the noun *porch* uttered by the speaker, the listener interprets the speaker's intended meaning correctly from context by choosing the most probable interpretation among a set of possible construals or paraphrases (bar length indicates probability of an interpretation). In comparison, the BERT language model assigns higher probabilities to inappropriate interpretations of the same denominal utterance. (b) The Google Translate system (evaluated in June 2021) incorrectly interprets *to porch the newspaper* as *to confuse the newspaper* when translating the query denominal utterance into Mandarin Chinese, which in turn leads to the erroneous back-translation into English.

idealized cognitive model (Lakoff 2008; Fauconnier 1997), and qualia (Pustejovsky 1991). In the context of noun-to-verb conversion, frame semantics theory provides a principled foundation for characterizing human interpretation and generation of novel denominal verb usages. For example, the utterance "the boy porched the newspaper" may be construed as invoking a NEWSPAPER DELIVERY frame that involves both explicit *frame elements* including the DELIVERER (the boy), the DELIVEREE (the newspaper), and the DESTINATION (the porch), as well as two latent elements that are left underspecified for the listener to infer: the main verb (also known as LEXICAL UNIT) that best paraphrases the action of the DELIVERER during the delivery event (e.g., *drop*), and the semantic relation between the DELIVERER and DELIVEREE (in this case can be described using a preposition "on/onto"). Interpreting novel denominal usages, therefore, can be considered as a task of implicit semantic constituents inference, which has been explored in the paraphrasing of other types of compound expressions such as noun-noun compounds (Shwartz and Dagan 2018; Butnariu et al. 2009), adjective-noun

pairing (Lapata 2001; Boleda et al. 2013), and logical metonymy (Lapata and Lascarides 2003).

The prevalence of denominal verb usages is not constrained to contemporary English. Apart from being observed in adult and child speech of different European languages (Clark 1982; Tribout 2012; Mateu 2001), denominal verbs also commonly appear in more analytic languages (i.e., languages that rely primarily on helper words instead of morphological inflections to convey word relationships) such as Mandarin Chinese, where the absence of inflectional morphemes allows highly flexible shift from one word class to another (Dongmei 2001; Fang and Shenghuan 2000; Si 1996). From a historical point of view, many denominal verbs have emerged after the established usages of their parent nouns. For instance, according to the Oxford English Dictionary, the word *advocate* had been exclusively used as a noun denoting "a person who recommends/supports something" before 1500s. However, this word grew a verb sense of "to act as an advocate for something" which later became popular so quickly that Benjamin Franklin in 1789 complained to Noah Webster about such an "awkward and abominable" denominal use (Franklin 1789). It is therefore constructive to consider and evaluate a general formal framework for noun-to-verb conversion that supports denominal verb inference and generation across languages and over time.

In this work, we develop and evaluate a probabilistic framework, *Noun2Verb*, to model noun-to-verb conversion couched in the tradition of frame semantics. Our work extends the previous study (Yu, El Sanyoura, and Xu 2020), which offers a probabilistic generative approach to model the meaning of denominal verb usages as a collection of frame elements. As illustrated in Figure 2a, we use a probabilistic graphical model to capture the dependency over denominal utterances and their underlying frame

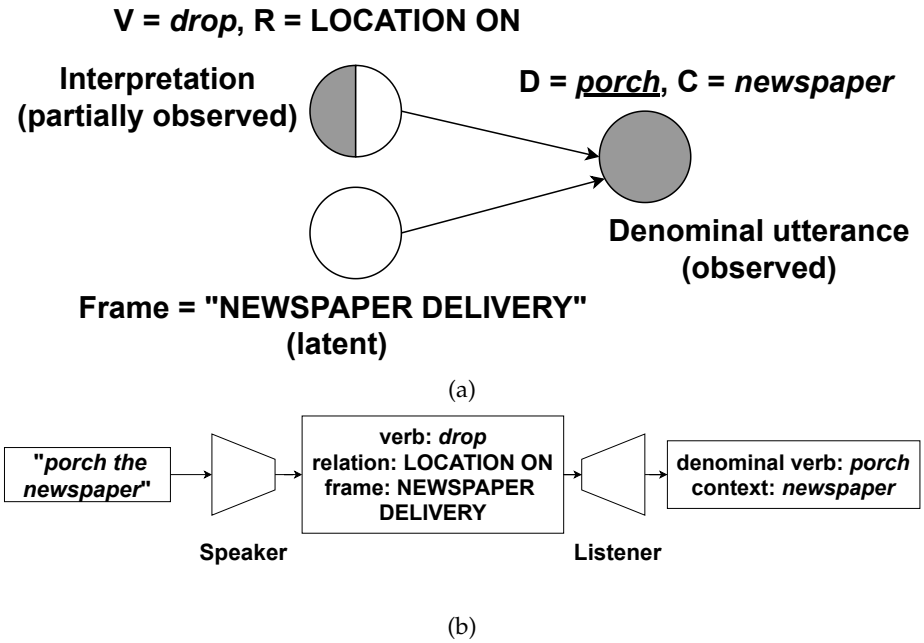


Figure 2
(a) The probabilistic graphical model of our *Noun2Verb* framework. (b) An illustration of the learning paradigm of *Noun2Verb* based on the reconstruction process.

elements, including (1) a partially observed interpretation of the denominal utterance consisting of a paraphrase verb and a semantic relation, and (2) a set of latent frame elements that further specify the underlying scenario. As shown in Figure 2b, our framework maximizes the joint probability of the three types of variables via a communicative process between a listener module and a speaker module. These modules learn collaboratively to reconstruct a novel denominal utterance. In particular, the listener would first observe an utterance with novel denominal usages, and "thinks out loud" about its appropriate interpretation, which is then taken by the speaker as a clue to infer the actual denominal utterance. Intuitively, this process can succeed only if the listener interprets the denominal utterance correctly, and the speaker shares similar semantic frame knowledge with the listener. This learning scheme therefore operationalizes the mutual-knowledge-based communication proposed in Clark and Clark (1979). Moreover, the reconstruction process also allows the models to learn from denominal utterances without explicit interpretation in an unsupervised way. To enable efficient learning, our framework draws on recent development from deep generative modeling (Kingma and Welling 2014; Kingma et al. 2014) and utilizes variational inference for training and learning with a minimal amount of labeled data.

Our current study extends earlier work showing how this probabilistic generative model provides automated interpretation and generation of novel denominal verb usages in modern English (Yu, El Sanyoura, and Xu 2020). We take a frame-semantic approach and compare three models of incremental complexity that range from a discriminative transformer-based model to a full generative model. We show that the transformer-based model, despite its success in many natural language understanding tasks (Devlin et al. 2018), is insufficient to capture the flexibility of denominal verbs and fails to productively generate novel denominal usages with relatively sparse training samples. We go beyond the previous work with a comprehensive evaluation of the framework with two additional sources of data: historical data of English noun-to-verb conversions and Mandarin denominal verb usages. Furthermore, we perform an in-depth analysis to interpret the learning outcomes of the generative model.

The remainder of this article is organized as follows. We first provide an overview of the relevant literature. We then present our computational framework, *Noun2Verb*, and specify the predictive tasks for model evaluation. We next present the datasets that we have collected and made publicly available for model learning and evaluation. We describe three case studies where we evaluate our framework rigorously on a wide range of data in contemporary English, Mandarin Chinese, and the historical development of English over the past two centuries. We finally provide detailed interpretations and discussion about the strengths and limitations of our framework and conclude.

2. Related Work

2.1 Computational Studies on Word Class Conversion

Compared to the extensive empirical and theoretical research on word class conversion, very few studies have attempted to explore this problem from a computational perspective. One of the existing studies leverages recent advances in distributed word representations and deep contextualized language models to investigate the directionality of word class conversion. In particular, Kisselew et al. (2016) build a computational model to study the factors that may account for historical ordering between noun-to-verb conversions and verb-to-noun conversions in English. In that study, they train a logistic regression model using bag-of-words embeddings of lemmas attested in both

nominal and verbal contexts to predict which word class (between noun and verb classes) might have emerged earlier in history. Their results suggest that denominal verbs usually have lower corpus frequencies than their parent noun counterparts, and nouns converted from verbs tend to have more semantically specific linguistic contexts. In a related recent study, Li et al. (2020) perform a computational investigation on word class flexibility in 37 languages by using the BERT deep contextualized language model to quantify semantic shift between word classes. They find greater semantic variation when flexible lemmas (i.e., lemmas that have more than one grammatical class) are used in their dominant word class, supporting the view that word class flexibility is a directional process.

Differing from both of these studies, here we focus on modeling the process of noun-to-verb conversion as opposed to the directionality or typology of word class conversion across languages.

2.2 Frame Semantics

The computational framework we propose is grounded in frame semantics, which has a long tradition in linguistics and computational linguistics. According to Fillmore, Johnson, and Petruck (2003), a semantic frame can potentially be evoked by a set of associated lexical units, which are often instantiated as the main predicate verbs in natural utterances. Each frame in the lexicon also enumerates several roles corresponding to facets of the scenario represented by the frame, where some roles can be omitted or null-instantiated and left underspecified for the listener to infer (Ruppenhofer and Michaelis 2014). The problem of interpreting denominal verb usages can therefore be considered as inferring (the concepts evoked by) latent lexical unit(s) of the underlying semantic frame, which is itself related to the tasks of semantic frame identification (Hermann et al. 2014) and semantic role labeling (Gildea and Jurafsky 2002). Given the limited available resources for labeled or fully annotated data, many existing studies have considered a generative and semi-supervised learning approach to combine annotated lexical databases such as FrameNet (Baker, Fillmore, and Lowe 1998) and PropBank (Kingsbury and Palmer 2002) with other unannotated linguistic corpora. For instance, the SEMAFOR parser presented by Das et al. (2014) is a latent variable model that learns to maximize the conditional probabilities of labeled semantic roles in FrameNet, and supports lexical expansion to unseen lexical units via the graph-based semi-supervised learning technique (Bengio, Delalleau, and Le Roux 2010). In a separate work, Thompson, Levy, and Manning (2003) learn a generative Hidden Markov Model using the labeled sentences in FrameNet and show that the resulting model is able to infer null-instantiated semantic roles in unobserved utterances (e.g., inferring that a "driver" role is missing given the sentence *The ore was boated down the river*).

Our framework builds on these existing studies by formulating noun-to-verb conversion as probabilistic inference of latent semantic frame constituents, and we suggest how a semi-supervised generative learning approach offers data efficiency and effective generalizations on the interpretation and generation of novel denominal verb usages that do not appear in the training data.

2.3 Models of Compound Paraphrasing

Our study also relates to a recent line of research on compound understanding. Many problems concerning the understanding of compounds require the inference of latent semantic constituents from linguistic context. For example, Nakov and Hearst (2006)

suggest that the semantics of a noun-noun compound can be expressed as multiple prepositional and verbal paraphrases (e.g., *apple cake* can be interpreted as *cake made of/contains apples*). Later work develops both supervised and unsupervised learning approaches to tackling noun-compound paraphrasing (Van de Cruys, Afantenos, and Muller 2013; Xavier and de Lima 2014). In particular, Shwartz and Dagan (2018) propose a semi-supervised learning framework for inferring the latent semantic relations of noun-noun compounds. They represent compounds and their paraphrases in a distributed semantic space parameterized by a biLSTM (Graves and Schmidhuber 2005) encoder. When paraphrases are not available, the missing components are replaced by the corresponding hidden representations yielded by the encoder. Shwartz and Dagan (2018) show good generalizability of their model on unobserved examples. We show that our framework generalizes well on novel denominal utterances due to a semi-supervised learning approach in a distributed semantic space, and further, the proposed framework can learn interpretation (listener) and generation (speaker) model simultaneously via generative modeling.

Previous linguistic studies also suggest that the lexical information in converted denominal verbs can be inferred from the listeners' knowledge about the intended referent of nominal bases (Baeskow 2006). It is therefore natural to connect noun-to-verb conversion to the linguistic phenomenon of logical metonymy, where language users need to infer missing predicates from certain syntactic constructions (e.g., *an easy book* means *a book that is easy to read*) (Pustejovsky 1991). Following this line of thought, Lapata and Lascarides (2003) propose a probabilistic model that can rank interpretations of given metonymical compounds by searching in a large corpus for their paraphrases, which are identified by exploiting the consistent correspondences between surface syntactic cues and meaning. We apply similar methods to extract candidate paraphrases of denominal utterances to construct our learning or training dataset, and we show that this frequency-based ranking scheme aligns reliably with human feasibility judgment of interpretations for denominal verb usages.

2.4 Deep Generative Models for Natural Language Processing

The recent surge of deep generative models has led to the development of several flexible language generation systems, such as variational autoencoders (VAEs) (Bowman et al. 2016; Bao et al. 2019; Fang et al. 2019) and generative adversarial networks (GANs) (Subramanian et al. 2017; Press et al. 2017; Lin et al. 2017). Our Noun2Verb framework builds on the architecture of semi-supervised VAE proposed by Kingma et al. (2014), where an interpretation/listener module and a generation/speaker module jointly learn a probability distribution over all denominal utterances and any of their available paraphrases. One advantage of VAEs is the ability to encode through their latent variables certain aspects of semantic information (e.g., writing style, topic, or high-level syntactic features), and to generate proper samples from the learned hidden semantic space via ancestral sampling. We show in our model analysis that the learned latent variables in our framework indeed capture the variation in both syntactic structures and semantic frame information of target denominal utterances and their paraphrases.

2.5 Deep Contextualized Language Models

For a sequence of natural language tokens, deep contextualized models compute a sequence of context-sensitive embeddings for each token. Many state-of-the-art natural language processing models are built upon stacked layers of a neural module

called the Transformer (Vaswani et al. 2017), such as BERT (Devlin et al. 2018), GPT-2 (Radford et al. 2019), RoBERTa (Liu et al. 2019), and BART (Lewis et al. 2020). These large neural network models are often pre-trained on predicting missing tokens given contextual information within a sentence. The models are then fine-tuned on learning examples of a series of downstream tasks including language generation tasks such as summarization, and natural language understanding tasks such as recognizing textual entailment. A common issue of most current transformer-based models is that many of their successful applications tend to rely on extensive fine-tuning on adopted benchmarks with (sometimes hundreds of) thousands of examples. For tasks where large-scale annotations of learning examples are infeasible, or where the target linguistic data are severely under-represented in standard pre-training resources, transformer models often yield much worse performance (Croce, Castellucci, and Basili 2020).

In our work, we consider a BERT-based language generation model as a competitive baseline, and we demonstrate that this pre-trained language model is insufficient to capture the flexibility of noun-to-verb conversions, particularly when ground-truth paraphrases for a denominal utterance are highly uncertain.

3. Computational Framework

We formalize noun-to-verb conversion as a dual problem of comprehension and production and formulate this problem under a frame semantics perspective. We present three incremental probabilistic models under differing assumptions about the computational mechanisms of noun-to-verb conversion.

3.1 Noun-to-verb Conversion as Probabilistic Inference

We consider noun-to-verb conversion as communication between a listener and a speaker over an utterance that includes a novel denominal verb usage. Our framework focuses on modeling the knowledge and dynamics that enable (1) a listener module to properly interpret the meaning of a novel denominal verb usage (or zero-shot inference) by paraphrasing, and (2) a speaker module to produce a novel denominal usage given an interpretation.

Figure 3 illustrates our framework. Here the speaker generates an utterance $U = (D, C)$ that consists of an innovative usage of denominal verb D (e.g., *porch*) and its *context* C. As an initial step, we consider the simple case where C is a single word that serves as the direct object of D (e.g., *newspaper* as the context for *porch*).

Given this utterance, the listener interprets its meaning M, which we operationalize as three key components: (1) a paraphrase verb V (e.g., *drop*) for the target denominal verb; (2) a *semantic relation* R following Clark and Clark (1979) that specifies the relation between the paraphrase verb and the context (e.g., an on-type location, signifying the fact that newspaper is dropped onto the porch); and (3) a set of frame elements E following the frame semantics tradition, which we elaborate below. The paraphrase verb V is an established verb that best describes the action denoted by the denominal verb D. It serves as the *lexical unit* that invokes the underlying semantic frame of D. The semantic relation R, according to empirical studies in Clark and Clark (1979), reflects how the novel sense of a denominal verb is extended from its parent noun, and falls systematically into eight main types (see a summary in Table 1). Within each relation type, there is a set of words (most of which are prepositions) that signify such a relation, along with a template paraphrase for denominal usages of this type. For instance, denominal usages of the form "to <denominal verb> the <context>" (e.g., "to porch

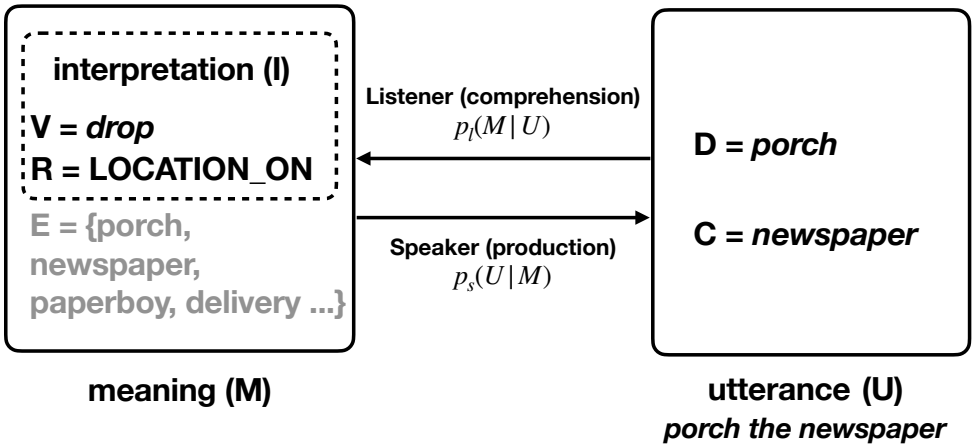


Figure 3
An illustration of the *Noun2Verb* framework. The speaker produces an utterance of a denominal verb usage from its production likelihood p_s. The listener interprets the meaning of the utterance by paraphrasing via its comprehension likelihood p_l. $U = (D, C)$ is the denominal utterance, where D is the target denominal verb, and C its object context; M is the meaning of U, where V is the paraphrased verb, R is the semantic relation, and E denotes a set of latent frame elements.

Table 1
Major types of semantic relation described in Clark and Clark (1979) that explain common denominal verb usages in English. Each semantic relation is specified by a set of relational words (mostly prepositions), and a syntactic schema that serves as the template for paraphrasing query denominal verb usages under a relation type.

Relation type	Relational words	Denominal usage	Template paraphrase
LOCATUM ON	*on, onto, in, into, to, at*	carpet the floor	put the carpet on the floor
LOCATUM OUT	*out (of), from, of*	shell the peanuts	remove the shell from the peanuts
LOCATION IN	*on, onto, in, into, to, at*	porch the newspaper	drop the newspaper on the porch
LOCATION OUT	*out (of), from, of*	mine the gold	dig the gold out of the mine
DURATION	*during*	weekend at the cabin	stay in the cabin during the weekend
AGENT	*as, like*	referee the game	watch the game as a referee
GOAL	*become, look like, to be, into*	orphan the children	make the children become orphans
INSTRUMENT	*with, by, using, via, through*	bike to school	go to school by bike

the newspaper") where D comes from the relation type LOCATION ON can usually be paraphrased as "to <paraphrase verb> the <context> onto/into/to the <denominal verb>" (e.g., "to drop the newspaper onto the porch"). Under a semantic frame invoked by V, the listener would simultaneously infer frame elements E that may be involved in the scenario expressed by the target utterance U—such inference captures not only participants that are explicitly specified by the denominal utterance, but also the residual contextual knowledge shared between the speaker and the listener that is not captured in variables V and R. In particular, *porch the newspaper* may invoke a DELIVERY frame, where one can identify that the element of DELIVERY is *the newspaper*, the destination is *the porch*, and infer that a reasonable choice of the DELIVERER role can be *the postman* or *the paperboy*. We denote $I = (V, R)$ as an interpretation for a target utterance U, while we specify frame elements E as latent variables (i.e., implicit knowledge) to be inferred by the models.

Our formulation drawing on semantic relations is motivated by existing cross-linguistic studies of denominal verbs. For example, Clark found that semantic relation types in Table 1 apply to many innovative denominal usages coined by children speaking English, French, and German (Clark 1982). A more recent comparative study of denominal verbs in English and Mandarin Chinese also found that these major semantic relations can explain many Chinese denominal verb usages (Bai 2014). The modeling framework we present here can automatically learn these semantic relations and latent frame elements from data, and, importantly, it can generalize to interpret and generate novel denominal usages across different languages and over time.

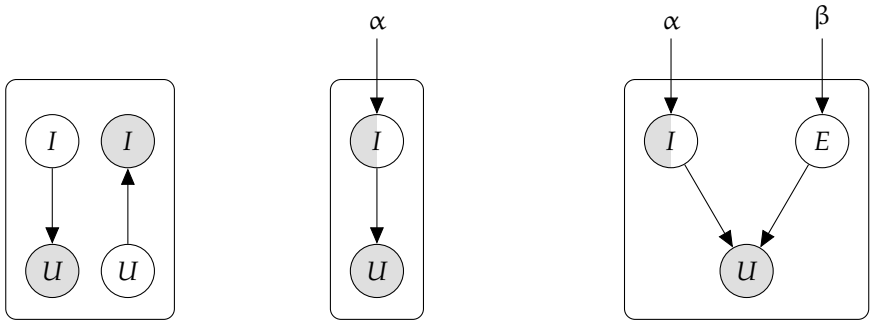
With the core components defined, we now formally cast noun-to-verb conversion as two related probabilistic inference problems. The listener module tackles the *comprehension problem* where given an utterance U, it samples appropriate paraphrases to interpret its meaning $M = (I, E) = (V, R, E)$ under the comprehension model $\mathrm{p}_l(M|U)$. The speaker module tackles the inverse *production problem* by producing a (novel) denominal usage U given an intended meaning M, under the production model $\mathrm{p}_s(U|M)$.

We postulate that mutually shared knowledge, when modeled as semantic frame information, should be key to successful communication for innovative denominal usages. To verify this view, we describe and examine three incremental probabilistic models under our framework dubbed *Noun2Verb*.

3.2 Model Classes

We present three probabilistic models (see illustrations in Figure 4) that make different assumptions about the computational mechanisms of noun-to-verb conversion. First, we describe a *discriminative model* that assumes neither any interactive dynamics between the speaker and the listener (i.e., no collaboration) nor any knowledge of semantic frame elements. We implement this model using a state-of-the-art contextualized language model from natural language processing. To our knowledge, there exists no specific and scalable model of denominal verbs, and, given the general-purpose nature of contextual language models, we consider it as a strong competitive baseline model. Next, we describe a *partial generative model* that enables listener-speaker collaboration via knowledge sharing but without any representation of semantic frame elements. Finally, we describe a *full generative model* that incorporates both listener-speaker collaboration and semantic frame elements.

3.2.1 Discriminative Model. The discriminative model consists of two sub-modules that learn separately without any collaboration or information sharing; hence it is



(a) The discriminative model (b) The partial-generative model (c) The full-generative model

Figure 4
A graphical illustration of the three probabilistic models under the proposed framework. U, I, and E stand for the variables of (u)tterance that contains a denominal verb usage, (i)ntended meaning of the utterance, and (e)lements of the semantic frame invoked by the utterance, respectively. α and β represent the hyperparameters for the prior distributions of the variables. Shaded, half-shaded, and unshaded circles represent latent variables, semi-latent variables, and observables.

insensitive to frame elements E in its meaning representation. As illustrated in Figure 4a, the listener module receives a denominal utterance, and produces a paraphrase of that utterance by sampling an interpretation from the conditional distribution $\mathrm{p}_l(I = (V, R)|U = (D, C))$. The speaker module reverses the listener module by generating a denominal utterance from the conditional distribution $\mathrm{p}_s(U = (D, C)|I = (V, R))$. During learning, we present the model with a supervised set (i.e., fully labeled data) of denominal utterances $X_\mathrm{s} = \{(U^{(i)}, I^{(i)})\}_{i=1}^{M}$. Each such data point is paired with a human-annotated ground-truth paraphrase verb and semantic relation (i.e., the interpretation for a query denominal usage). We optimize the speaker-listener modules by minimizing the standard negative log-likelihood classification loss for both modules independently:

$$\mathcal{S} = \mathcal{S}_l + \mathcal{S}_s \tag{1}$$

$$\mathcal{S}_l = -\sum_{(U^{(i)}, I^{(i)}) \in X_\mathrm{s}} \log \mathrm{p}_l(I^{(i)}|U^{(i)}; \Theta_l) \tag{2}$$

$$\mathcal{S}_s = -\sum_{(U^{(i)}, I^{(i)}) \in X_\mathrm{s}} \log \mathrm{p}_s(U^{(i)}|I^{(i)}; \Theta_s)) \tag{3}$$

Here Θ_s and Θ_l denote the parameters under the speaker and the listener modules, respectively. This discriminative learner bears resemblance to compound phrase understanding systems, where classification models are trained to predict implicit semantic relations that hold between phrase constituents (Shwartz and Dagan 2018).

We consider the state-of-the-art language model BERT (Devlin et al. 2018) to parameterize the listener and speaker distributions (see Appendix A for a detailed description of the BERT model architecture). However, as we will demonstrate empirically, despite the incorporation of such a powerful neural language model with a rich knowledge base, this discriminative baseline model is insufficient to simulate word class conversion

in sensible ways, mostly due to its limitations in capturing the flexibility and uncertainty involved in natural denominal usages. For instance, both "drop the newspaper on the porch" and "leave the newspaper on the porch" can be considered good interpretations for the query denominal usage *porch the newspaper*, but systems like BERT, as shown later, tend to idiosyncratically favor a very restricted set of construals and cannot account for the fine-grained distribution of human interpretations for denominal usages. Furthermore, the speaker and listener modules in the discriminative model do not share mutual knowledge by jointly encoding the same probability distribution over denominal utterances and their interpretations; that is, $\mathrm{p}_l(I|U)$ and $\mathrm{p}_s(U|I)$ do not necessarily induce the same joint distribution $\mathrm{p}(U, I)$. We therefore turn to a more cognitively viable generative model by incorporating the interaction between the listener and speaker modules to encourage agreement on the utterance-meaning distributions—a prerequisite for successful communication with innovative denominal usages (Clark and Clark 1979).

3.2.2 Partial Generative Model. The partial generative model, illustrated in Figure 4b, defines a generative process of how a speaker might produce a novel denominal usage. We first sample an interpretation by drawing I from a categorical prior distribution $\mathrm{p}_0(I|\alpha)$ parametrized by α. We then feed this interpretation to the speaker module so as to sample a novel denominal utterance via $\mathrm{p}_s(U|I)$. This setup enforces a joint utterance-interpretation distribution $\mathrm{p}_s(U, I|\alpha) = \mathrm{p}_0(I|\alpha)\mathrm{p}_s(U|I)$, which allows us to operationalize the idea of shared mutual knowledge by encouraging the listener to be consistent with the speaker when interpreting novel denominal usages.

Formally, we learn the listener's likelihood $\mathrm{p}_l(I|U)$ as a good approximation for the speaker's distribution $\mathrm{p}_s(I|U)$ over interpretations:

$$\mathrm{p}_\mathrm{l}(I|U) \approx \mathrm{p}_\mathrm{s}(I|U) \tag{4}$$

We parametrize model distributions p_l and p_s via feed-forward neural networks (see Appendix A for a detailed model description). One advantage of this generative approach is that it supports learning with sparse labeled data. In particular, this model can learn from a handful of labeled data and an unlabeled, unsupervised set $X_\mathrm{u}\{(U^{(i)})\}_{i=1}^{N}$, where each denominal verb usage has no human annotation (in terms of its meaning).

To learn this model, we apply an optimization technique known as variational inference, commonly used for generating data with highly complex structures, including images (Narayanaswamy et al. 2017) and text (Semeniuta, Severyn, and Barth 2017), to train the two modules simultaneously. Let Θ again denote the set of all parameters in the model; we optimize Θ by minimizing the following evidence lower bound (ELBO) loss function:

$$\mathcal{U} = \sum_{U^{(i)} \in X_\mathrm{u}} \mathbb{E}_{I \sim \mathrm{p}_\mathrm{l}}[\log \mathrm{p}_\mathrm{s}(U|I)] - D[\mathrm{p}_\mathrm{l}(I|U)||\mathrm{p}_0(I|\alpha)] \tag{5}$$

Here $\mathbb{E}_{I \sim p_\mathrm{l}}(\cdot)$ refers to taking the expectation by sampling interpretation I from the listener's conditional likelihood $p_\mathrm{l}(I|U)$, and $D(\cdot||\cdot)$ denotes the Kullback-Leibler (KL) divergence between two probability distributions. This learning scheme does not require any labeled interpretation I. Instead, the two modules learn collaboratively by seeking to reconstruct a denominal verb utterance: The first term $\mathbb{E}_{I \sim p_\mathrm{l}}[\log p_\mathrm{s}(U|I)]$ of

$\mathcal{U}$ describes a scenario where the listener first observes a U and "thinks out loud" about its interpretation I, which is then taken by the speaker (who is hidden from the utterance) as a clue to infer the actual utterance. Intuitively, if the listener understands U reasonably, and provided that the speaker shares a similar utterance-interpretation mapping with the listener, the reconstruction is more likely to succeed, and existing theoretical analyses validate this idea (see, for example, Rigollet [2007], for detailed discussion). It can be shown that minimizing $\mathcal{U}$ is equivalent to maximizing the joint log-likelihood of all denominal utterances in the unsupervised set, while simultaneously finding a listener's likelihood $p_{\mathrm{l}}(I|U)$ that best approximates the speaker's posterior $p_{\mathrm{s}}(I|U)$. We provide the proof of this equivalence in Appendix B for interested readers.

Apart from the above unsupervised learning procedure, we can also train the two modules separately on the labeled, supervised set X_s just as we learn in the discriminative model. The overall learning objective $\mathcal{L}$, therefore, consists of minimizing jointly a supervised loss term and an unsupervised one, which can be operationalized through the paradigm of semi-supervised learning:

$$\mathcal{L} = \mathcal{U} + \lambda\mathcal{S} \tag{6}$$

Here $\mathcal{S}, \mathcal{U}$ are the two losses defined in Equations (1) and (5), and λ is a hyperparameter controlling the relative weighting of the supervised and unsupervised data. Training the partial generative model (as well as the full generative model described next) is algorithmically equivalent to learning a semi-supervised variational autoencoder proposed by Kingma et al. (2014).

3.2.3 Full Generative Model. Similar to the partial model, the full generative model illustrated in Figure 4c also defines a generative process from meaning M to utterance D, except that the semantic frame elements E are incorporated as a latent variable: $\mathrm{p}_s(U, I|\alpha, \beta) = \mathrm{p}_0(I|\alpha)\mathrm{p}_0(E|\beta)\mathrm{p}_s(U|I)$, where α, β are hyperparameters that define the categorical priors of I and E, respectively. In this model, both the interpretation I and the semantic frame E give rise to a denominal utterance. Intuitively, the introduction of frame elements helps the model to further distinguish denominal utterances of similar interpretations but distinct intended referents. For example, both of the denominal utterances (1) *carpet the floor* and (2) *blanket the bed* can be paraphrased by the same coarse, semantic-relation template "to put A on (the top of) some B", but their actual contexts are quite different. The frame element E is expected to capture such fine-grained variation in meaning by learning the residual contextual information underspecified by V and R. Similar to the partial generative model, we still expect an agreement between the posteriors of meaning $\mathrm{p}_s(M|U)$ and $\mathrm{p}_l(M|U)$, but here we use the full representation of $M = (I, E)$ by taking frame elements into consideration:

$$\mathrm{p}_{\mathrm{l}}(I, E|U) \approx \mathrm{p}_{\mathrm{s}}(I, E|U) \tag{7}$$

The listener and speaker distributions here are parametrized by neural network encoders. During learning, the model can also be trained via a mixture of (1) reconstruction of unlabeled denominal utterance, and (2) inference and generation of labeled denominal usages with ground-truth paraphrases. The unsupervised learning stage is also

conducted through variational inference with an ELBO loss function similar to the partial model:

$$\mathcal{U} = \sum_{U^{(i)} \in X_u} \mathbb{E}_{(I,E)\sim p_l}[\log p_s(U|I,E)] - D[p_l(I,E|U)||p_0(I|\alpha)p_0(E|\beta)] \tag{8}$$

whereas the supervised learning loss is identical to L in Equation (1), and the overall semi-supervised loss shares the same form as specified in Equation (4).

3.3 Specification of Predictive Tasks

We consider our models in two predictive tasks: (1) in the *comprehension task*, the listener module of the model takes an utterance containing a novel query denominal usage U and provides an interpretation of its meaning through sampling from $p_l(I|U)$ it defines; and (2) in the *production task*, the speaker module conversely generates a novel denominal usage U from its $p_s(U|I)$ based on a query meaning specified in I. For the full generative model, since U depends on both interpretations and frame elements, we apply a Monte Carlo approach to approximate $p_s(U|I)$ and $p_l(I|U)$ by first drawing a set of frame elements $E^{(k)}$ from model priors, and then taking the average over the production probabilities $p_s(U|I,E^{(k)})$ induced by sampled elements:

$$p_l(I|U) \approx \sum_{E^{(k)}\sim p_0(E|\alpha)} p_l(I,E^{(k)}|U) \tag{9}$$

$$p_s(U|I) \approx \sum_{E^{(k)}\sim p_0(E|\alpha)} p_s(U|I,E^{(k)}) \tag{10}$$

For evaluation against historical data, we incrementally predict the denominal usages $U^{(t+\Delta)}$ of a target noun D emerged at future time $t+\Delta$, given its established noun usages up to time t—for instance, we expect the model to infer whether the noun "phone" can grow out a verb sense given its nominal usage before 1880s. We formalize this temporal prediction problem by assuming that an appropriate denominal usage generated by the speaker should be acceptable to the language community in the future. We thus extend the synchronic production task to make diachronic prediction. In particular, the speaker module takes the predicate verbs and semantic relations associated with the target noun D at time t as interpretation $I^{(t)}$, and sample a denominal usage $\hat{U}^{(t)} \sim p_s(U|I^{(t)})$ as model prediction for denominal usages into the future times $t+\Delta$:

$$\Pr(U^{(t+\Delta)}|I^{(t)}) = p_s(U^{(t)}|I^{(t)}) \tag{11}$$

The full generative model $p_s(U^{(t)}|I^{(t)})$ is again approximated by the Monte Carlo sampling approach in Equation (8).

4. Data

To evaluate our framework comprehensively against natural denominal verb usages, we collected three datasets: (1) denominal verbs from adults and children speaking contemporary English extracted from the literature (DENOM-ENG); (2) denominal verbs in contemporary Mandarin Chinese extracted from the literature (DENOM-CHN); and

(3) denominal verbs extracted from historical English corpora (DENOM-HIST). Each dataset consists of a supervised set X_s of denominal usages with interpretations, and an unsupervised set X_u of unannotated denominal usages. We also collected a set of synchronic denominal verb usages with ground-truth paraphrases annotated via online crowdsourcing (DENOM-AMT) for model evaluation.[1] The experimental protocol of this work has been approved by the research ethics boards at the University of Toronto (REB # 00036310). A total amount of 1,304 US dollars were paid to human annotators for about 13,000 responses. Every annotator received an estimated hourly payment that is higher than the minimum wage requirement in their registered country.[2]

4.1 Denominal Verb Usages from English-speaking Adults and Children (DENOM-ENG)

Clark and Clark (1979) provide a large list of denominal verb utterances (i.e., a denominal verb with its context word) from English adults, and Clark (1982) also reports a set of novel denominal uses produced by English-speaking children under age 7. Although all of these denominal utterances are labeled with their ground-truth relation types R, none of them has ground-truth paraphrase verb(s) V available. To obtain natural interpretations of denominal meaning (for constructing the supervised set for model learning), we searched for the top 3 verbs that co-occur most frequently with each denominal utterance using the paraphrase templates specified in Table 1 (and we validated these searched results using crowdsourcing described later). We performed these searches in the large-scale comprehensive iWeb 2015 corpus (`https://corpus.byu.edu/iweb/`), specifically through the Sketch Engine online corpus tool (`https://www.sketchengine.eu`) and its built-in regular-expression queries—for example, a denominal utterance "to carpet the floor" with a "LOCATUM ON" relation type would have a paraphrase utterance template "to __ the carpet on/onto the floor", where "__" is filled by a verb. We obtained 786 annotated denominal utterances from adult data, and 32 annotated examples from children.

While a small portion of denominal utterances has explicit human-annotated paraphrases, a greater proportion does not have such information. We expect our models to be able to interpret novel denominal verb usages by generalizing from the small set of annotated data and also learning from the large set of unlabeled data. For example, if the model is told that "send the resume via email" is the correct paraphrase for *email the resume*, then on hearing a similar utterance like *mail the package*, it should generalize and infer that utterance has something to do with the transportation frame (as in the case with *mail*). To facilitate such "frame borrowing" learning, we obtained a set of novel denominal usages by replacing the denominal verb D of each U described previously with a semantically related noun (e.g., *mail the letter* → *email the letter*). We took the taxonomy from WordNet (`https://wordnet.princeton.edu/`) and extracted all synonyms of each denominal verb D from the same synset as substitutes. This yielded 1,129 novel utterances examples for unsupervised learning.

1 Data and code for our analyses are available at the following repository: `https://github.com/jadeleiyu/noun2verb`.

2 The average payment in our task is 33.6 USD per hour, which is above the minimum wage requirements of the registered countries of all involved participants (from Canada, People's Republic of China, United Kingdom, and United States).

4.2 Denominal Verb Usages in Mandarin Chinese (DENOM-CHN)

Similar to the case of English, noun-to-verb conversion has been extensively investigated in Mandarin Chinese. In particular, Bai (2014) performed a comparative study of denominal verbs in English and Mandarin Chinese by collecting over 200 examples of noun-to-verb conversions in contemporary Chinese, and categorizing these denominal usages under the same relation types described by Clark and Clark (1979). It was found that the eight major relation types of English denominal verbs can explain most of their Chinese counterparts, despite some small differences. We therefore extend our probabilistic framework of English noun-to-verb conversion to model how Chinese speakers might comprehend and produce denominal verb usages, hence testing the generality of our proposed framework to represent denominal meaning in two very different languages.

Similar to DENOM-ENG, we performed an online corpus search on the iWeb-2015-Chinese corpus via Sketch Engine to determine the top 3 most common paraphrase verbs for each Chinese denominal utterance. This frequency-based corpus search yields a supervised set of 230 Chinese denominal utterances. We also augmented DENOM-CHN by replacing the denominal verb D of each U in Bai (2014) with a set of synonyms taken from the taxonomy of Chinese Open WordNet database (Wang and Bond 2013). After excluding cases with morphological or tonic changes during noun-to-verb conversions, we obtained an unsupervised set of 235 denominal utterances.

4.3 Denominal Verb Usages in Historical Development of English (DENOM-HIST)

To determine English nouns that had a temporal noun-to-verb conversion in history, we used the syntactically parsed Google Books Ngram Corpus that contains the frequency of short phrases of text (*ngrams*) from books written over the past two centuries (Goldberg and Orwant 2013).

We first extracted time series of yearly counts for words (*1-grams*) whose numbers of occurrence as nouns and verbs both exceed a frequency threshold θ_f, and we computed the proportion of noun counts for each word w as follows:

$$Q(w,t) = \frac{\#(w \text{ as a noun at year } t)}{\#(w \text{ as a noun at year } t) + \#(w \text{ as a verb at year } t)} \tag{12}$$

We then applied the change-point detection algorithm introduced by Kulkarni et al. (2015) to find words with a statistically significant shift in noun-to-verb part-of-speech (POS) tag ratio. This method works by detecting language change over a general stochastic drift and accounting for this by normalizing the POS time series. The method identifies change points via bootstrapping under a null hypothesis that, in most cases, the expected value of a word's POS percentage should remain unchanged (compared to random fluctuations). Therefore, by permuting the normalized POS time series, the pivot points with the highest shifts in mean percentage would be the statistically significant change points. Applying this method yielded a set of 57 target words as denominal verbs for our diachronic analysis. Since the n-gram phrases in Google Syntactic-Ngrams (GSN) are too short (with maximum length of 5 words) to extract complete denominal utterances and paraphrases, we considered another historical English corpus, the Corpus of Historical American English (COHA), which comprises annotated English sentences from the 1810s to 2000s. We assumed that each denominal verb w has been exclusively used as a noun prior to $t^*(w)$, and we extracted paraphrase

usages $I^{(t)}$ before $t^*(w)$ as conventional usages, and denominal utterances $U^{(t)}$ after $t^*(w)$ as novel usages for prediction. All denominal utterances and paraphrases with aligned contextual objects and targets are taken as the supervised set, while the denominal utterances without aligned paraphrases found in the historical corpus are used for unsupervised learning, yielding an X_s of size 1,055 and an X_u of size 8,972.

4.4 Crowd-sourced Annotation of Denominal Verb Usages (DENOM-AMT)

We evaluate our models on a set of denominal utterances with high-quality ground-truth paraphrases interpreted by human annotators. We collected human interpretations for a subset of the English and Chinese denominal verb usages in the training set described above via Amazon Mechanical Turk (AMT) crowdsourcing platform.

For each utterance D, we presented the online participants with the top 3 paraphrase verbs collected from the iWeb corpora via frequency-based search, and asked the participants to choose, among the 3 candidates, all verbs that serve as good paraphrases for the target denominal verb in the denominal utterance. If none of them is appropriate, then the participants must provide a good alternative paraphrase verb by themselves. All annotators of English and Mandarin Chinese denominal verb examples must have passed a qualification test to confirm their proficiency in the respective languages to participate in the annotation.[3] This online crowdsourcing procedure yields 744 annotated examples in English and 55 examples in Chinese (24 English utterances in DENOM-ENG were discarded due to insufficient number of collected responses).[4] For each utterance in English, there are on average 14.7 responses and 2.43 unique types of paraphrase verbs collected, while for Chinese we obtain 12.8 responses and 1.97 paraphrase verb types per utterance. The resulting dataset includes in total 606 unique types of English denominal verbs and 54 unique types of Chinese denominal verbs. The English annotators reached an agreement score of $\kappa = 0.672$ measured with Cohen's Kappa, and $\kappa = 0.713$ for Chinese annotators. For English questions, 407 out of 744 denominal utterances have at least one alternative paraphrase provided in the annotations; for Chinese questions, 19 out of 55 utterances have at least one alternative paraphrase.

5. Evaluation and Results

We first describe the experimental details and the procedures for evaluation of our proposed framework. We then present three case studies that evaluate this framework against different sources of innovative denominal usages drawn from speakers of different age groups and across languages, as well as data from contemporary and historical periods.

3 See `https://github.com/jadeleiyu/noun2verb/tree/main/data/annotations` for questionnaires of language proficiency test and denominal utterance interpretation.

4 We did not collect human responses for all examples in DENOM-CHN because many denominal uses have become obsolete in contemporary Mandarin (though they still appear in formal text such as official documents and therefore can be found via web corpus search). The first author therefore manually selected 54 Chinese denominal verbs considered to have nominal meanings familiar to modern Mandarin speakers.

5.1 Details of Experimentation and Evaluation

We ran the proposed probabilistic models on the 3 training datasets (DENOM-ENG, DENOM-CHN, and DENOM-HIST) by optimizing over their loss functions specified in Section 2. The speaker and listener modules in partial and full generative models are implemented as three-layer feed-forward neural networks using the Pyro deep probabilistic programming library (Bingham et al. 2019). For the discriminative model in the contemporary datasets, we initialized both the listener and speaker modules with 12-layer pre-trained BERT neural language models implemented by the HuggingFace library based on PyTorch, and we fine-tuned the parameters in BERT during training. The input sequences (I, E for listener modules, and U for speaker modules) were first encoded via distributed word embeddings, which were then fed into the corresponding modules for further computation. For synchronic prediction, we applied the GloVe algorithm (Pennington, Socher, and Manning 2014) on Wikipedia 2014 and Gigaword 5 corpora to learn distributed word embeddings.

To initialize the models and prevent these embeddings from smuggling in information about target denominal verb usages for model prediction, we removed all denominal usages for target denominal verbs during training. For historical prediction, we replaced GloVe embeddings with the HistWords historical word embeddings used in Hamilton, Leskovec, and Jurafsky (2016) for each decade from the 1800s to 1990s. Similar to the synchronic case, we re-trained all historical embeddings by explicitly removing all denominal usages of each target word D (that we seek to predict) from the original text corpora.[5]

We assess each model on the evaluation set of denominal verb usages that have ground-truth paraphrases, in the two types of predictive tasks described. In the comprehension tasks, for each novel denominal utterance U, we sample interpretation from the listener module's posterior distribution $p_l(I|U)$, and compare these model predictions against the ground-truth paraphrases provided by human annotators. In the production tasks, we conversely group all denominal utterances that share the same verb-relation pair as ground-truth interpretations of the intended meaning (e.g., "mail my resume" and "email my number" with common paraphrase verb "send" and relation "INSTRUMENT"). For every interpretation, we apply the speaker module to generate novel denominal usages from the posterior distribution $p_s(U|I)$.

We consider two metrics to evaluate model performance: (1) Standard receiver operating curves (ROCs), which provide a comprehensive evaluation for the model prediction accuracy based on $k = 1, 2, 3, ...$ guesses of interpretation/utterances from its posteriors. Prediction accuracy (or precision) is the proportion of interpretations/utterances produced by the model that fall into the set of ground-truths—this metric automatically accounts for model complexity and penalizes any model that has poor generalization or predictive ability; we also report the mean accuracy when considering only the top-k model predictions from $k = 1$ up to $k = 5$. (2) Kullback-Leibler divergence D_{KL}, on the other hand, measures the ability of the models to capture fine-grained human annotations. Because each query denominal verb usage has multiple ground-truths (i.e., the set of paraphrases provided by human annotators that form a

5 We validated the reliability of the POS tagger by asking human annotators on AMT to manually inspect 100 randomly sampled denominal utterances detected by SpaCy from the iWeb-2015 corpus. We collected 5 responses for each utterance, and found that at least 3 annotators agree with the automatically labeled POS tags for 94 out of 100 cases.

ground-truth distribution), we compute the discrepancy between the empirical distributions $p_{\text{emp}}(U|I)$, $p_{\text{emp}}(I|U)$ of paraphrases/utterances collected from AMT workers, and model-predicted posteriors $p_l(I|U)$ and $p_s(U|I)$—a smaller D_{KL} indicates better alignment between the natural distribution of human judgment and model posterior distribution (see Appendix D for details of calculating the KL divergence). Since the value of KL divergence may be sensitive to the size of the evaluation set, we calculate the D_{KL} for contemporary English examples by randomly sampling from DENOM-AMT 100 subsets of English denominal utterances with the same size of the Chinese evaluation set, and taking the mean KL divergence between the sampled sets of utterances and their ground-truth paraphrases.

For the case of contemporary English, because almost all supervised examples are human-annotated, we adopt a 12-fold leave-one-out cross-validation procedure: First, we randomly split the 744 annotated utterances into 12 equally sized subsets; second, we draw 11 subsets of annotated examples for model training (together with unsupervised utterances for generative models), and use the left out subset for model evaluation. We repeat this procedure 12 times and report the mean model performance. For the case of contemporary Chinese, we train models using all denominal utterances in DENOM-CHN (i.e., 235 unannotated utterances and 230 annotated examples with paraphrases determined via corpus search), and evaluate the models using the 55 human-annotated Chinese denominal utterances in DENOM-AMT. For the diachronic analysis in English, we keep a subset of the supervised learning example X_s as test cases (and remove them from training sets). We also consider two additional baseline models for evaluating if any of the three proposed probabilistic models can learn above chance, at all: (1) a frequency-based model that chooses the interpretation and denominal usage with highest empirical probability $p_{\text{emp}}(U|I)$ or $p_{\text{emp}}(I|U)$ in the training dataset, and (2) a random-guess model that samples each linguistic component from a uniform distribution.

5.2 Case Study 1: Contemporary English

We first evaluate the models in comprehension and production of contemporary English denominal verb usages. Figures 5a and 5c summarize the results on 744 English denominal utterances in DENOM-AMT using ROCs, while Figures 6a and 6c show the predictive accuracy on the same evaluation dataset when considering the top-k outputs for each model, with k ranging from 1 to 5.

We found that all non-baseline models achieved good accuracy in predicting semantic relation types (lowest accuracy 96%), so we focused our discussion on model predictions of interpreting via paraphrased verbs V and generating novel denominal verbs D. We computed the area-under-the-curve (AUC) statistics to compare the cumulative predictive accuracy of the models, summarized also in Figure 5. Our full generative model yields the best AUC scores and top-k predictive accuracy for both tasks, outperforming the partial generative model, which is in turn superior to the BERT-based discriminative model.

The left two columns of Table 2 show the mean KL-divergence scores between model posteriors and empirical distributions over both interpretations and denominal utterances on all 744 English test cases in DENOM-AMT. We observed that both full and partial generative models offer better flexibility in interpreting and generating novel denominal verb usages, but the discriminative model, despite its high predictive accuracy, yields output distributions that are least similar to human word choices among non-baseline models. In particular, we found that the generative models outperform

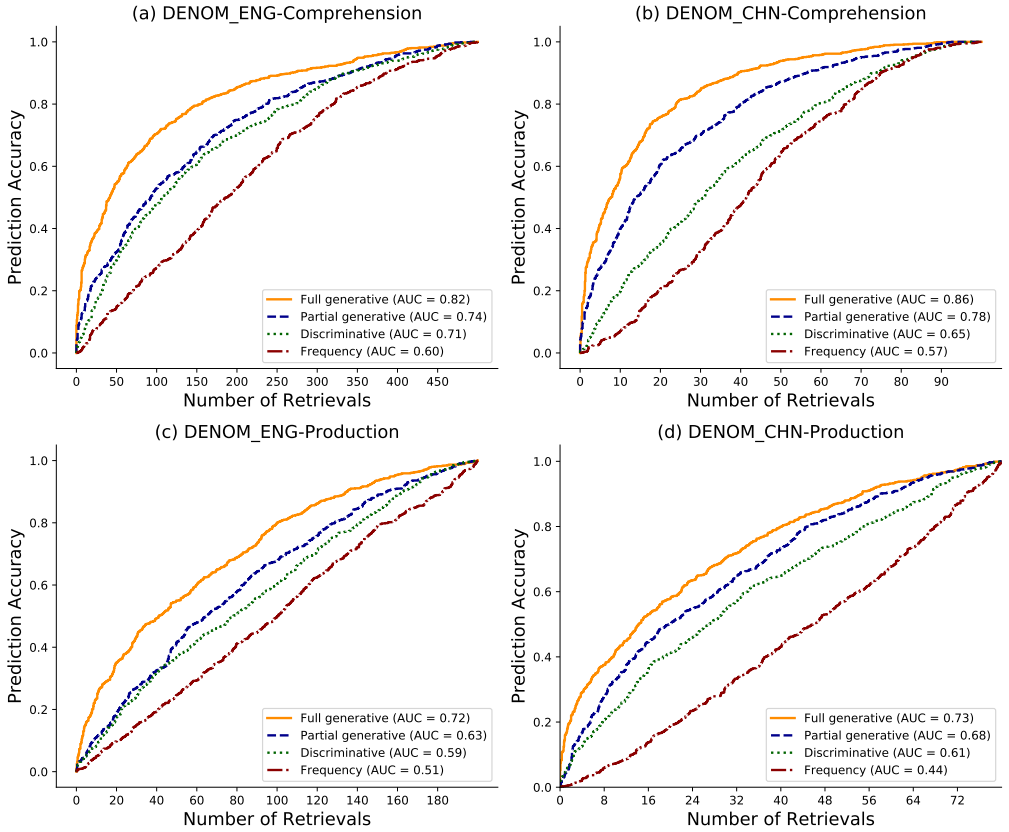


Figure 5
A summary of model performance in denominal verb comprehension and production. The left column summarizes the results from the 744 English examples in the DENOM-AMT dataset based on receiver operating characteristic (ROC) curves, and the right column summarizes similar results from the 55 Chinese examples in the DENOM-AMT dataset. "Frequency" refers to the frequency baseline model. Higher area-under-the-curve (AUC) score indicates better performance.

their discriminative counterparts on both child and adult denominal utterances (see Appendix C for a detailed breakdown of these results).

To demonstrate the better flexibility of the full generative model, we visualize in the first row of Figure 7 the listener posterior distributions $p_l(V|U)$ over paraphrase verbs for the query denominal usage $U =$"porch the newspaper" based on the full generative and discriminative models (top 20 candidates with non-zero probabilities are shown). We found that the full generative model assigned the highest posterior probabilities on the three ground-truth human-annotated verb paraphrases *dropped*, *left*, and *threw*, and the partial generative model also ranked them as the top 5 candidates (posterior of which is not shown in the figure). In contrast, the discriminative model only assigned the highest posterior probability for *drop*, and failed to distinguish the two alternative ground-truths between other implausible candidate paraphrase verbs. The second and third most likely candidates predicted by the discriminative model are *saw* and *wanted*, most possibly because these are commonly associated words in the pre-training text corpora of BERT. This limitation of not being able to explain the fine-grained

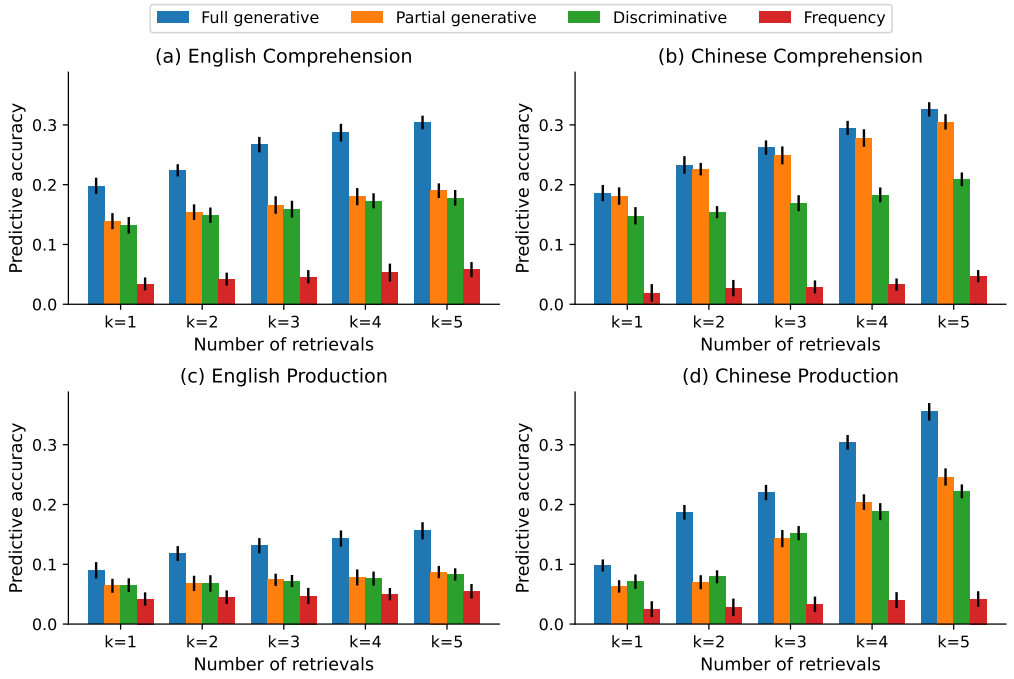


Figure 6
Model predictive accuracy in denominal verb comprehension and production when taking the top-k outputs, with k ranging from 1 to 5. The left column summarizes the results from the 744 English examples in the DENOM-AMT dataset, and the right column summarizes similar results from the 55 Chinese examples in DENOM-AMT. "Frequency" refers to the frequency baseline model. Vertical bars represent standard errors.

Table 2
Model comparison on predicting human annotated denominal data. Model accuracy is summarized by Kullback-Leibler (KL) divergence between posterior distributions $p_{\text{comp}}(V|U)$, $p_{\text{prod}}(D|I)$, and fine-grained empirical distributions of human-annotated ground-truth on the DENOM-AMT dataset. A lower value in KL indicates better alignment between model distribution and empirical distribution. Standard errors are shown within the parentheses.

Model	KL divergence ($\times 10^{-3}$)			
	English		Chinese	
	Comprehension	Production	Comprehension	Production
Full Generative	8.86 (1.1)	21.7 (2.4)	2.93 (0.46)	7.8 (1.1)
Partial Generative	10.01 (0.9)	22.4 (2.5)	3.08 (0.35)	11.0 (1.1)
Discriminative	13.75 (1.0)	39.0 (1.8)	3.32 (0.33)	29.4 (1.8)
Frequency Baseline	11.41 (0)	57.7 (0)	3.62 (0)	28.5 (0)

distribution of paraphrases and only locking onto a single best solution has appeared to be a general issue for the discriminative model, as we observed the same phenomenon in many other test cases.

We also examine whether our generative models outperform the baselines in the comprehension tasks by simply favoring paraphrase verbs that are more frequently

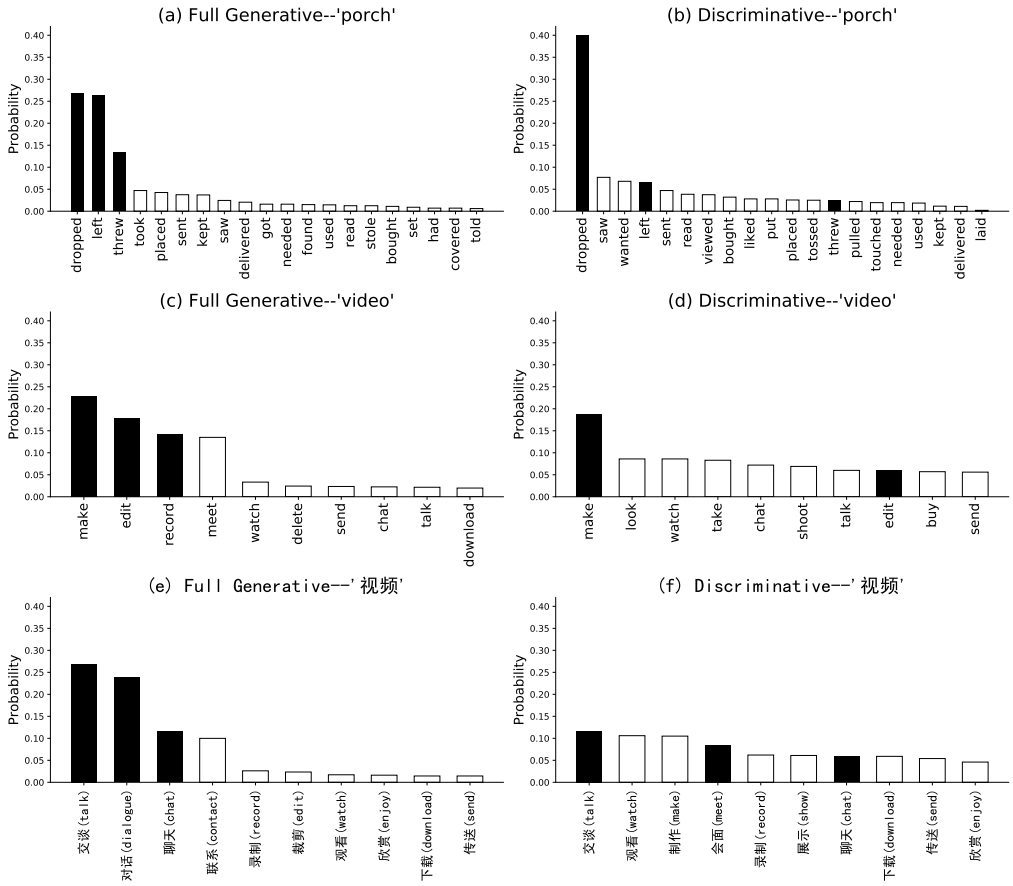


Figure 7
Comparison of the full generative model and the discriminative model on the quality of paraphrasing novel denominal usages. The top 2 rows show model posterior distributions on the paraphrase verbs (horizontal axis) for the query denominal verbs *porch* and *video* in English. The bottom row shows similar information for the query denominal verb *video-chat* in Chinese, with the English translations in parentheses. Model predicted probabilities for the top 3 choices from human annotation (i.e., ground truth) are shown in solid black bars.

paired with the target denominal verbs in the linguistic corpora. The left two columns of Table 3 summarize individual model predictive accuracy on interpreting denominal utterances in DENOM-AMT that have ground-truth paraphrases either completely generated by corpus search (i.e., no alternative paraphrases collected) or produced as alternative paraphrases by human annotators. For corpus-generated cases, we calculate the model comprehension accuracy when considering the top-3 output paraphrase verbs; for human-generated cases with m alternative interpretations, we calculate the model comprehension accuracy when considering the top-m output paraphrase verbs. We found that the full generative model offers most accurate interpretations in both cases, while the discriminative model, despite its high accuracy in predicting corpus-generated paraphrases, performs significantly worse on human-generated examples. These results further demonstrated the inflexibility of discriminative language models on interpreting denominal utterances, especially when the feasible paraphrase verbs rarely co-occur with a given target noun in the reference corpora.

Table 3
Model comparison on predicting human-generated and corpus-generated denominal utterance paraphrases.

Model	Comprehension accuracy			
	English		Chinese	
	Human	Corpus	Human	Corpus
Full Generative	0.406	0.459	0.279	0.288
Partial Generative	0.373	0.408	0.243	0.257
Discriminative	0.297	0.462	0.222	0.276
Frequency Baseline	0.153	0.307	0.097	0.109

These initial findings on contemporary English data therefore suggest a generative, frame-based approach to denominal verbs that encodes speaker-listener shared knowledge and latent semantic frames.

5.3 Case Study 2: Contemporary Chinese

We next investigate whether the same framework generalizes to denominal verb usages in a language that is markedly different from English. Table 4 presents some exemplar denominal verbs that are common in contemporary Mandarin Chinese. Although Chinese does not have morphological markings for word classes, there exist alternative linguistic features that signify the presence of a verb. For example, a word that appears after the auxiliary word "地" or before "得" is typically a verb. If a word that is commonly used as a noun appears in context of these verbal features, it can then be considered as a denominal verb. For example, the phrase "开心地__" denotes "to __ happily", where "__" is filled with a verb. Therefore, when a noun such as "视频" (video) appears in the phrase "开心地视频" (to video happily), a Chinese speaker would consider "视频" as a verb converted from its nominal class. It is worth noting that for some Chinese nouns, their direct translations into English are still valid denominal verbs, but their meaning may differ substantially in two languages. For instance, the denominal utterance "I videoed with my friends" would remind an English speaker of a scenario of video recording, while for Chinese speakers the correct interpretation should be "I chat with my friends via online video applications". We therefore expect our models to be able to capture such nuanced semantic variation when learning from the Chinese denominal dataset DENOM-CHN.

Table 4
Examples of denominal verb usages in contemporary Mandarin Chinese, together with their literal translations and interpretations in English. The target Chinese denominal verbs and their English translations are underlined, and their corresponding English paraphrase verbs are marked in bold font.

Chinese denominal verb usage	Literal translation in English	Interpretation in English
漆门窗	paint doors windows	to **paint** doors and windows
网鱼	net fish	to **catch** fish with the net
圈地	circle land	to **enclose** land
和朋友视频	with friends video	to **chat** with friends via webcam

The right columns of Figures 5 and 6 show, respectively, the ROC curves (with AUC scores) and the top-k predictive accuracy of each model on the comprehension and production tasks in Mandarin Chinese. We observed that, similar to the case of English denominal verbs, the full generative model yields the best overall predictive accuracy. Moreover, as shown in the right two columns of Table 2, the generative model aligns better with Chinese speakers' interpretation and production of denominal verbs, because it yields a lower KL-Divergence score between its posteriors and empirical distributions of ground-truths in comparison to the discriminative model. Moreover, as shown in the right two columns of Table 3, the performance of the discriminative still drops significantly when switching from predicting corpus-generated to human-generated paraphrases, while its generative counterparts are much less influenced by this change.

As an example of where our framework successfully captures cross-linguistic semantic variation in denominal verbs, the second and third rows of Figure 7 illustrate the posterior distributions over paraphrase verbs for the discriminative and full generative models on the exemplar utterance "to video (视频) with my friends", in English and Chinese. In both languages, the full generative model assigns the highest probability masses on the three ground-truth paraphrases chosen by human annotators, thus demonstrating its ability to flexibly interpret a denominal verb based on its linguistic context under different languages. The discriminative model not only favors idiosyncratically a single ground-truth verb ("交谈 [talk]" for Chinese and "make" for English), but it also yields less flexible model posteriors when translating from English to Chinese. For instance, the BERT model fails to realize that "make" should no longer be a good paraphrase in Chinese, while the full generative model successfully excludes this incorrect interpretation from the top 10 candidates in the listener's posterior distribution.

These results further support our generative, frame-based approach to denominal verbs that explains empirical data in two different languages.

5.4 Case Study 3: Historical English

In our final case study, we examine whether our framework can predict the emergence of novel denominal verb usages in the history of English.

We first demonstrate the effectiveness of our change point detection algorithm in determining valid noun-to-verb conversions in history. Figure 8 shows the word frequency-based Z-score series $Z(w)$ for the noun-to-verb count ratio series $Q(w, t)$ of some sample words, together with the change points $t^*(w)$ detected by our algorithm. We observed that the algorithm correctly identifies substantial shifts in noun-to-verb count ratios across time.

We next report the temporal predictive accuracy of our models by dividing the evaluation set of denominal verbs into groups where change points fall under the same decade. For each target word D with m novel denominal usages observed after its detected change point $t^*(w)$, we take the top m sampled usages with the highest speaker's posterior probability $p_s(U|I)$ as the set of retrieved model predictions, and we then calculate the average predictive accuracy of the top m predictions over all denominal verbs emerged in each decade. We considered two kinds of evaluation criteria when making predictions for target D in future decade t: taking as ground truth denominal utterances that contain D (1) specifically in the immediate following decade $t + 1$, and (2) in any future decade $t' > t$ up to the terminal decade 2000s.

Figure 9 shows the decade-by-decade precision accuracy for all the three probabilistic models, along with a frequency baseline that always chooses the top m denominal utterances with the highest frequencies that contain D. The predictive accuracy falls

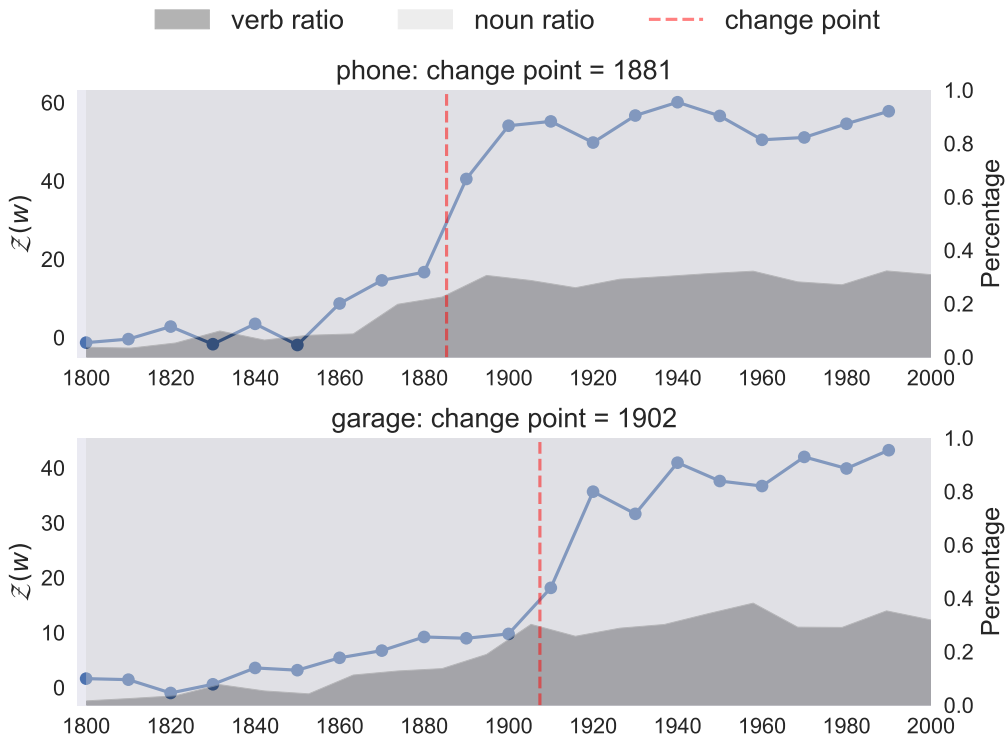


Figure 8
Frequency-based z-score time series $Z(w) = Z(Q(w,t))$ for the nouns "phone" and "garage" over the past two centuries. The stacked color areas denote percentage of noun/verb usage frequencies in each year, and the red vertical lines mark the detected historical change point of noun-to-verb conversion.

systematically in later decades because there are fewer novel denominal verbs to predict in the data. To ensure that most target nouns have attested denominal uses in future decades, we only calculate model precision up to 1980s. The full generative model yields consistently better results in almost every decade, and it is followed by the BERT-based discriminative model. Note that in later decades, the difference in accuracy of BERT model and the full model becomes smaller, presumably due to the increasing similarity between learning data and text corpora on which BERT is pre-trained (i.e., contemporary text corpora). Overall, our generative model yields more accurate prediction on denominal verb usages than the discriminative model with rich linguistic knowledge.

Taken together, these results provide firm evidence that our probabilistic framework has the explanatory power over the historical emergence of denominal verb usages.

6. Model Interpretation and Discussion

To interpret the full generative model, we visualize the latent frame representations learned by this model. We also discuss the strengths and limitations from qualitative analyses of example denominal verb usages interpreted and generated by the model.

Classic work by Clark and Clark (1979) provided a fine-grained classification of denominal usages within the relation type INSTRUMENT. We use this information to gain an intuitive understanding of the full generative model in its ability to capture fine-grained semantics via the frame element variable E. Figure 10 shows the t-distributed

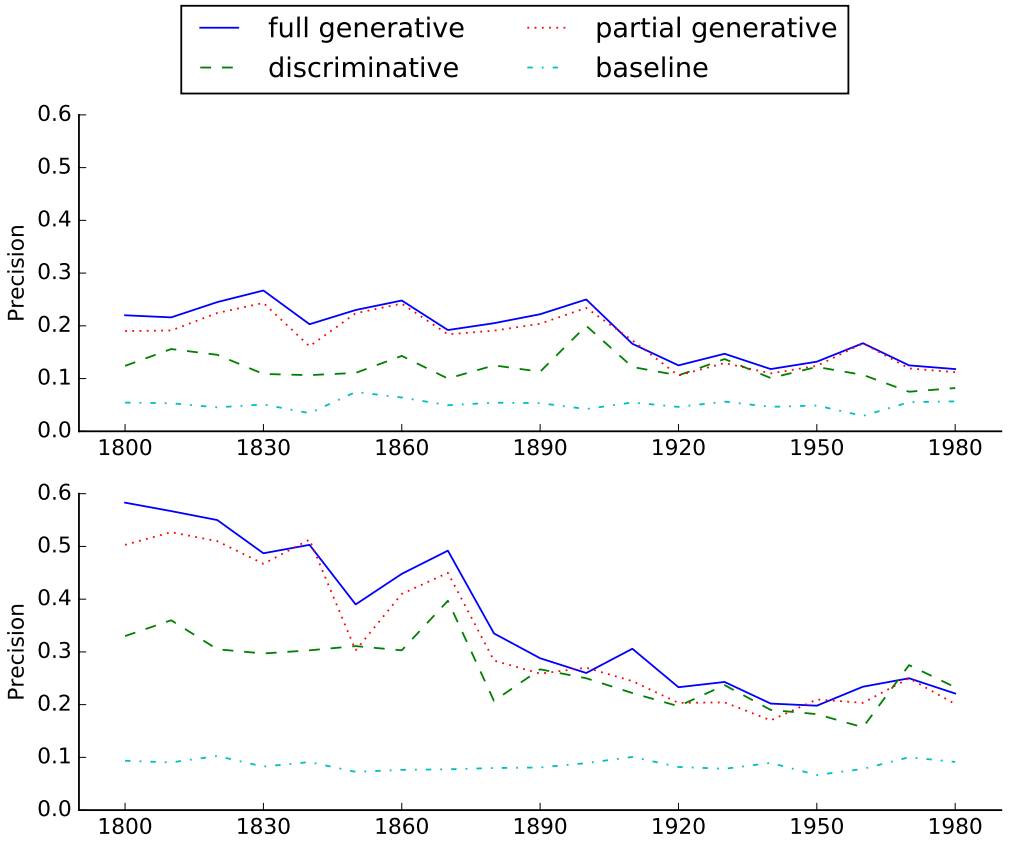


Figure 9
Model predictive accuracy on the DENOM-HIST dataset. Top row shows the predictive accuracy where only emerging denominal verb usages in the immediate next decade are considered. Bottom row shows predictive accuracy where all future emerging denominal usages are considered.

Stochastic Neighbor Embeddings (t-SNE) (van der Maaten and Hinton 2008), a nonlinear dimensionality reduction method that projects high-dimensional data into low-dimensional space for visualization, of the model-learned latent frame elements for the example denominal utterances in INSTRUMENT type. All data points are expressed in markers following their sub-category labels pre-specified in Clark and Clark (1979). We observe that the learned frame variable E encourages denominal usages within the same sub-category to be close in semantic space, manifested in the four clusters. As such, the frame variable helps to capture the fine-grained distinctions of denominal usages (even within the same broad semantic category).

To gain further insight into the generative model, we show in Tables 5 and 6 example model predictions in the three datasets we analyzed, for denominal verb comprehension and production, respectively. In the comprehension task (Table 5), our model provides reasonable interpretations on novel denominal verbs that did not appear in the learning phase. For instance, the model inferred that *blanket the bed* can be paraphrased as "put/drop/cover the blanket on the bed", which are close approximations to the top 3 ground-truth paraphrases "put/place/lay the blanket on the bed". One factor that facilitated such inference is that there are analogous denominal verbs during model

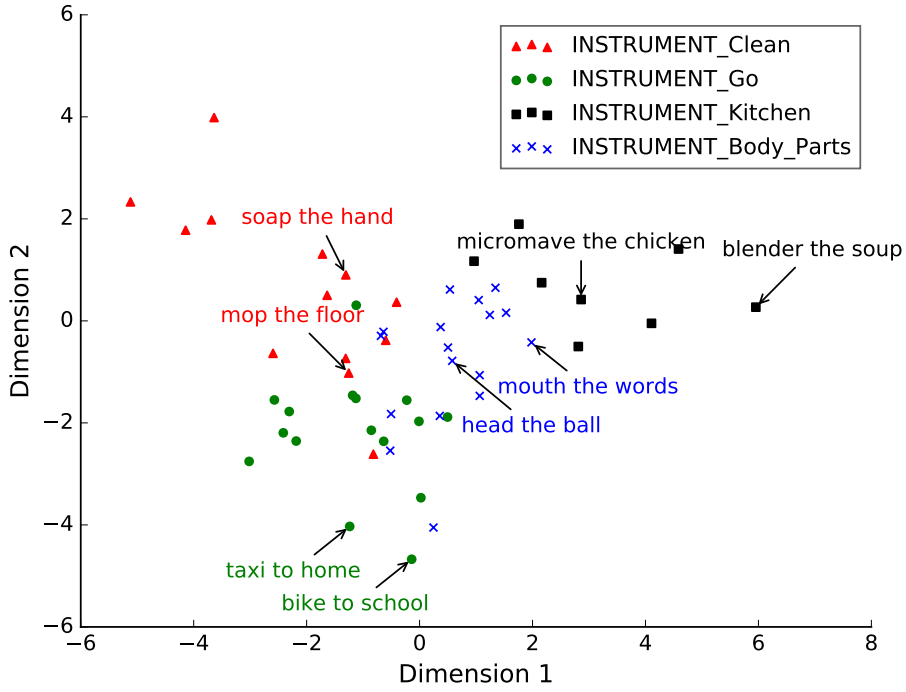


Figure 10
t-distributed Stochastic Neighbor Embedding visualization of the model-learned frame elements (E) for denominal verb usages in the category of INSTRUMENT (DENOM-ENG dataset). Utterances within each sub-category based on Clark and Clark (1979) are shown with the same marker. For each sub-category, the most frequent denominal utterances are annotated with its source text.

Table 5
Example paraphrases of novel denominal usages interpreted by the full generative model.

Denominal usage	Dataset	Semantic relation	Human paraphrases with model-predicted ranks in ()	Top verb paraphrases inferred from the model
blanket the bed	DENOM-ENG	locatum on	put(1), place(3), lay(11)	put, drop, cover
paper my hands	DENOM-ENG	instrument	cut(1), hurt(2)	cut, hurt, wound
fox the police	DENOM-ENG	agent	deceive(4), baffle(2), fool(3)	cheat, baffle, fool
网鱼 (net the fish)	DENOM-CHN	instrument	捕(catch, 1), 抓(capture, 3)	捕(catch), 捉(seize), 抓(capture)
garage the car	DENOM-HIST	location on	stop(1), put(6), park(2)	stop, park, move
mine the gold	DENOM-ENG	location out	dig(327), extract(609), get(25)	put, bury, find
bee the cereal	DENOM-ENG	locatum on	add(54)	get, find, eat

learning (e.g., *carpet the floor*) that allowed the model to extrapolate to new denominal cases.

In the production task (Table 6), our model successfully generated both (1) now conventionalized denominal usages such as *stem the flowers*, even though *stem* was completely missing in the learning data, and (2) truly novel denominal cases such as *chimpanzee my gestures*, presumably by generalization from training examples such as *parrot my words*.

Table 6
Examples of novel denominal usages produced by the full generative model.

Paraphrase verb	Dataset	Semantic relation	Ground-truth denominal verb utterances	Denominal usages sampled from model posterior
remove	DENOM-ENG	instrument	shell the peanuts, fin the fish, skin the rabbit	stem the flowers
repeat	DENOM-ENG	agent	parrot my words	chimpanzee my gestures
选(choose)	DENOM-CHN	instrument	筛人(sieve the candidates)	筛选书籍(sieve the books)
冷却(freeze)	DENOM-CHN	instrument	冰水(ice the water)	冰食物(ice the food)

Similar to the case of English, models trained on Chinese denominal data also exhibit generalizability. However, we also observed poor model generalization, especially when training instances from a semantic type are extremely sparse. For instance, as Clark and Clark (1979) point out, very few English denominal verbs fall under the "location out" relation type, and our model therefore often misinterpreted denominal usages under this type (e.g., *mine the gold*) as cases from "location in" type (e.g., "put/bury/find the gold in(to) the mine"). These failed cases suggest that richer and perhaps more explicit ontological knowledge encoded in denominal meaning, such as the fact that mines are places for excavating minerals, should be incorporated. These failed cases might also be attributed to the difficulty in distinguishing different types of semantic relations with word embeddings, such as synonyms from antonyms.

Another issue concerns the semantic representation of more complex denominal verbs that cannot be simply construed via paraphrasing, but are otherwise comprehensible for humans. For example, the denominal usage *bee the cereal* was observed in a sub-corpus of the CHILDES dataset, where a child asked the mother to add honey to his cereal. Interpreting such an innovative utterance requires complex (e.g., a chain of) reasoning by first associating bees with honey, and then further linking honey and cereal. Because the paraphrase templates we worked with do not allow explanations such as "to add the honey (produced by bees) into the cereal", all models failed to provide reasonable interpretations for this novel usage.

Our framework is not designed to block novel denominal usages. Previous studies suggest that certain denominal verbs are not productive due to blocking or statistical preemption, for example, we rarely say *car to work* because the denominal use of *car* is presumably blocked by the established verb *drive* (Clark and Clark 1979; Goldberg 2011). We believe that this ability of knowing what not to say cannot be acquired without extensive knowledge of the linguistic conventions of a language community, although we do not consider this aspect as an apparent weakness of our modeling framework (since *car to work* does have sensible interpretations for English speakers, even though it is not a conventional expression in English). In contrast, we think it is desirable for our framework to be able to interpret and generate such preemptive cases, because such expressions though blocked in one language can be productive and comprehensible in other languages. Our generative models are able to produce and interpret such potential overgeneralizations due to their exposure to unsupervised denominal utterances generated from synonym substitutions as explained in Section 4.1.

7. Conclusion

We have presented a formal computational framework for word class conversion with a focus on denominal verbs. We formulate noun-to-verb conversion as a dual

comprehension and production problem between a listener and a speaker, with shared knowledge represented in latent semantic frames. We show in an incremental set of probabilistic models that a generative model encoding a full distribution over semantic frames best explained denominal usages in contemporary English (by adults and children), Mandarin Chinese, and the historical development of English. Our results confirmed the premise that probabilistic generative models, when combined with structured frame semantic knowledge, can capture the comprehension and production of denominal verb usages better than discriminative language models. Future work can explore the generality of our approach toward characterizing other forms of word class conversion. Our current study lays the foundation for developing natural language processing systems toward human-like lexical creativity.

Appendix A. Design Details of the Neural Network Modules

Recall that the listener's and the speaker's distributions of three probabilistic models are parametrized by deep neural networks. For discriminative models, we use the BERT transformer to compute hidden representations for each token of the input sequence, and pass them through a fully connected layer (parametrized by a transformation matrix W and a bias vector b) to obtain proper probability distributions:

$$\mathrm{p}_l(I = (V, R)|U) = \sigma(W_{l,V} \cdot f_{\mathrm{BERT}}(U) + b_{l,V}) \cdot \sigma(W_{l,R} \cdot f_{\mathrm{BERT}}(U) + b_{l,R}) \tag{A.1}$$

$$\mathrm{p}_s(U = (D, C)|I) = \sigma(W_{s,D} \cdot f_{\mathrm{BERT}}(I) + b_{s,D}) \cdot \sigma(W_{s,C} \cdot f_{\mathrm{BERT}}(I) + b_{s,C}) \tag{A.2}$$

Here σ is the softmax function, and we assume that, conditioned on the input sequence, each component of the output sequence can be generated independently.

Similar to the discriminative model, both modules in the partial generative model first map input sequence into a hidden semantic space, and then sample each token of the output sequence independently by computing a categorical distribution for each component:

$$\mathrm{p}_l(I = (V, R)|U) = \sigma(W_{l,V} \cdot f_{\mathrm{l}}(U) + b_{l,V}) \cdot \sigma(W_{l,R} \cdot f_{\mathrm{l}}(U) + b_{l,R}) \tag{A.3}$$

$$\mathrm{p}_s(U = (D, C)|I) = \sigma(W_{s,D} \cdot f_{\mathrm{s}}(I) + b_{s,D}) \cdot \sigma(W_{s,C} \cdot f_{\mathrm{s}}(I) + b_{s,C}) \tag{A.4}$$

For the full generative model, the listener would also sample frame elements E during inference, and the speaker would also take frame elements E as input when sampling denominal utterances during generation:

$$\mathrm{p}_l(I, E|U) = \sigma(W_{l,V} \cdot f_{\mathrm{l}}(U) + b_{l,V}) \cdot \sigma(W_{l,R} \cdot f_{\mathrm{l}}(U) + b_{l,R}) \cdot \sigma(W_{l,E} \cdot f_{\mathrm{l}}(U) + b_{l,E}) \tag{A.5}$$

$$\mathrm{p}_s(U|I, E) = \sigma(W_{s,D} \cdot f_{\mathrm{s}}(I, E) + b_{s,D}) \cdot \sigma(W_{s,C} \cdot f_{\mathrm{s}}(I, E) + b_{s,C}) \tag{A.6}$$

Appendix B. Mathematical Proofs for Variational Learning

Here we show that the variational learning scheme described can achieve the following two goals simultaneously: (1) finding a probabilistic model that maximizes likelihood of all unsupervised denominal utterances, and (2) finding a pair of listener-speaker modules that induce the same joint distribution $\Pr(U, I)$ over denominal utterances and their interpretations. We shall use the full generative model for the proof, although the results should also apply to the partial generative model.

Suppose we have an unsupervised set of denominal utterances without any paraphrases available. To compute the probability that our generative model would generate a particular utterance U, we need to consider each possible meaning M that may be associated with it, and then sum up all joint probabilities $\mathrm{p}_s(U, M)$ defined by the model:

$$\mathrm{p}_s(U) = \sum_M \mathrm{p}_s(U, M) \tag{B.1}$$

The log-likelihood $\mathcal{J}$ of all utterances therefore has the form:

$$\mathcal{J} = \sum_{U^{(i)} \in X_\mathrm{u}} \log[\sum_M p_\mathrm{s}(U^{(i)}|M) p_0(M)] \tag{B.2}$$

$$= \sum_{U^{(i)} \in X_\mathrm{u}} \log[\mathbb{E}_{M \sim p_0}[p_\mathrm{s}(U^{(i)}|M)]] \tag{B.3}$$

where we use $\mathbb{E}_{M \sim p_0}$ to denote the process of taking expectation over all possible meanings. However, optimizing $\mathcal{J}$ directly would be difficult for most cases, and a common alternative is first finding a lower bound of $\mathcal{J}$, and then maximizing that bound—this is where we introduce the listener into the learning process. In particular, by inserting a listener's posterior $p_l(M|U)$ as an approximation of speaker's belief in utterance meanings $p_s(M|U)$, we can re-write Equation (18) as:

$$\mathcal{J} = \sum_{U^{(i)} \in X_\mathrm{u}} \log[\sum_M p_\mathrm{s}(U^{(i)}|M) p_0(M)] \tag{B.4}$$

$$= \sum_{U^{(i)} \in X_\mathrm{u}} \log[\sum_M \frac{p_\mathrm{s}(U^{(i)}|M) p_0(M)}{p_l(M|U)} p_l(M|U)] \tag{B.5}$$

where we divide the joint probability $p_s(U, M)$ with the listener's meaning likelihood $p_l(U|M)$ and multiply it back. Using Jensen's Inequality and the concavity of the log function, we can therefore derive a lower bound for $\mathcal{J}$ by replacing the log-of-sum in Equation (20) with a sum-of-log term:

$$\mathcal{J} = \sum_{U^{(i)} \in X_\mathrm{u}} \log[\sum_M \frac{p_\mathrm{s}(U^{(i)}|M) p_0(M)}{p_l(M|U)} p_l(M|U)] \tag{B.6}$$

$$\geq \sum_{U^{(i)} \in X_\mathrm{u}} \sum_M p_l(M|U) \log \frac{p_\mathrm{s}(U^{(i)}|M) p_0(M)}{p_l(M|U)} \tag{B.7}$$

$$= \sum_{U^{(i)} \in X_\mathrm{u}} \sum_M [\log p_s(U|M) - \log \frac{p_l(M|U)}{p_0(M)} p_l(M|U)] \tag{B.8}$$

$$= \sum_{U^{(i)} \in X_\mathrm{u}} \mathbb{E}_{M \sim p_0}[\log p_s(U|M)] - D[p_\mathrm{l}(M|U)||p_0(M|\alpha)] \tag{B.9}$$

$$= \mathcal{U} \tag{B.10}$$

Therefore, the unsupervised loss defined in Equation (3) is a universal lower bound of $\mathcal{J}$. Ideally, $\mathcal{U}$ and $\mathcal{J}$ would converge after training. In this case, the log-of-sum term in Equation (21) will be equal to the sum-of-log in Equation (22), implying that the fractional term $\frac{p_s(U^{(i)}|M)p_0(M)}{p_l(M|U)}$ inside the log becomes constant (i.e., independent of M):

$$\frac{p_s(U^{(i)}|M)p_0(M)}{p_l(M|U)} = c \tag{B.11}$$

Moving $p_l(M|U)$ to the right-hand side and integrating over M we have:

$$p_s(U^{(i)}) = c \tag{B.12}$$

So p_l becomes the following:

$$p_l(M|U) = \frac{p_s(U^{(i)}|M)p_0(M)}{c} \tag{B.13}$$

$$= \frac{p_s(U^{(i)}|M)p_0(M)}{p_s(U^{(i)})} \tag{B.14}$$

$$= p_s(M|U) \tag{B.15}$$

via Bayes' rule. Therefore, by optimizing $\mathcal{U}$, we not only maximize the log-likelihood J of all denominal utterances, but also operationalize the idea of shared semantic knowledge by forcing the listener and speaker module to define the same joint utterance-meaning distribution.

Appendix C. Prediction of Denominal Verb Usages in English-speaking Adults and Children

Table C1
Model comparison on predicting human annotated English denominal utterances made by adults and children. Model accuracy is summarized by Kullback-Leibler (KL) divergence between posterior distributions $p_{\text{comp}}(V|U)$, $p_{\text{prod}}(D|I)$, and fine-grained empirical distributions of human-annotated ground-truth on DENOM-AMT dataset. A lower value in KL indicates better alignment between model distribution and empirical distribution. Standard errors are shown within the parentheses.

Model	KL divergence ($\times 10^{-3}$)			
	English adults		English children	
	Comprehension	Production	Comprehension	Production
Full Generative	16.8 (2.4)	53.1 (4.6)	29.7 (3.0)	92.5 (1.4)
Partial Generative	19.1 (1.7)	56.5 (5.5)	31.1 (3.5)	115.7 (1.4)
Discriminative	34.7 (2.2)	103.1 (3.9)	30.6 (2.9)	104.6 (1.3)
Frequency Baseline	44.7 (0)	133.2 (0)	44.7 (0)	133.2 (0)

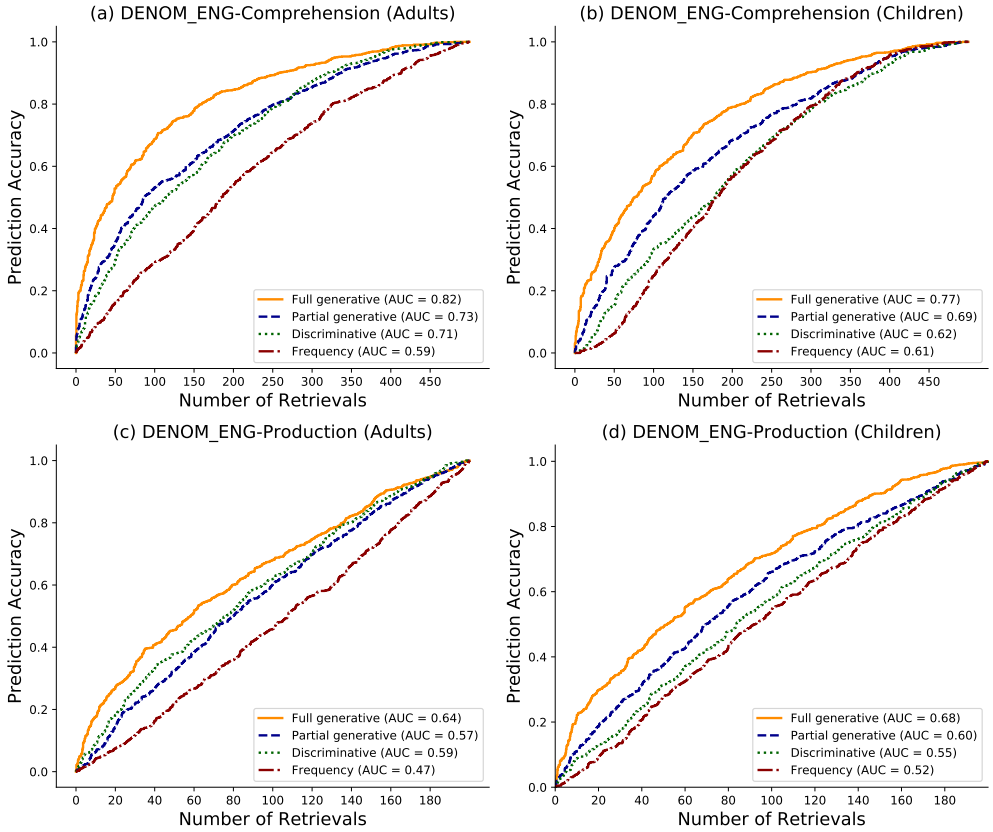


Figure C1
A breakdown of model performance in English denominal verb comprehension and production, based on adults' and children's usage data. The left column summarizes the results from the 100 denominal utterances made by adults in the DENOM-AMT dataset based on receiver operating characteristic (ROC) curves, and the right column summarizes similar results from the denominal utterance in DENOM-AMT made by children. "Frequency" refers to the frequency baseline model. Higher area-under-the-curve (AUC) score indicates better performance.

Appendix D. Calculation of KL Divergence

The Kullback-Leibler (KL) divergence D_{KL} measures how one probability distribution is different from another reference distribution. For two discrete distributions P, Q, their KL divergence is defined as:

$$D_{KL}(P||Q) = \sum_{x} P(x) \log(\frac{P(x)}{Q(x)}) \tag{D.1}$$

In our analysis, we compute the KL divergence between (1) the distribution of ground-truth responses in DENOM-AMT (P, proportional to the number of votes returned by human annotators), and (2) the model's output distribution of predicted

word (Q). For instance, consider the following question of paraphrasing a denominal utterance:

"I carpet the floor." → "I___the carpet on the floor." (D.2)

Suppose that there are 4 candidate paraphrase verbs with non-zero votes by the annotators and non-zero output probabilities returned by the full generative model:

Candidate verb	put	drop	place	leave
Vote by annotators	8	2	5	1
Induced empirical probability (P)	0.5	0.13	0.31	0.06
Output probability by full generative model (Q)	0.41	0.08	0.16	0.01

The KL divergence can therefore be calculated as the following:

$$D_{KL}(P||Q) = 0.5 * \log\frac{0.5}{0.41} + 0.13 * \log\frac{0.13}{0.08} + 0.31 * \log\frac{0.31}{0.16} + 0.02 * \log\frac{0.02}{0.01} = 0.38 \quad \text{(D.3)}$$

Acknowledgments

We thank Graeme Hirst and Peter Turney for helpful feedback on our manuscript. We thank Lana El Sanyoura, Bai Li, and Suzanne Stevenson for constructive comments on an earlier draft of our work. We also thank Aotao Xu, Emmy Liu, and Zhewei Sun for helping with the experimental design. This research is supported by a NSERC Discovery Grant RGPIN-2018-05872349, a SSHRC Insight Grant #435190272, and an Ontario Early Researcher Award #ER19-15-050 to YX.

References

Baeskow, Heike. 2006. Reflections on noun-to-verb conversion in English. *Zeitschrift für Sprachwissenschaft*, 25(2):205–237. https://doi.org/10.1515/ZFS.2006.008

Bai, Rong. 2014. Denominal verbs in English and Mandarin from a cognitive perspective. Master's thesis, University of Malaya.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90. https://doi.org/10.3115/980845.980860

Bao, Yu, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xin-yu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019. https://doi.org/10.18653/v1/P19-1602

Bengio, Yoshua, Olivier Delalleau, and Nicolas Le Roux. 2010. Label propagation and quadratic criterion. In *Semi-Supervised Learning*. MIT Press, chapter 11.

Bingham, Eli, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. 2019. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 20:1–6.

Boleda, Gemma, Marco Baroni, The Nghia Pham, and Louise McNally. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 35–46.

Bowman, Samuel R., Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on*

Computational Natural Language Learning, pages 10–21. https://doi.org/10.18653/v1/K16-1002

Butnariu, Cristina, Su Nam Kim, Preslav Nakov, Diarmuid O Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. Semeval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 100–105. https://doi.org/10.3115/1621969.1621987

Clark, Eve. 1982. The young word maker: A case study of innovation in the child's lexicon. *Language Acquisition: The State of the Art*, pages 390–425.

Clark, Eve and H. H. Clark. 1979. When nouns surface as verbs. *Language*, 55:767–811. https://doi.org/10.2307/412745

Croce, Danilo, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119. https://doi.org/10.18653/v1/2020.acl-main.191

Das, Dipanjan, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56. https://doi.org/10.1162/COLI_a_00163

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Dirven, René. 1999. Conversion as a conceptual metonymy of event schemata. *Metonymy in Language and Thought*, 275:287. https://doi.org/10.1075/hcp.4.16dir

Dongmei, Wang. 2001. Dissertation of denominal verbs of modern Chinese from cognitive view. *Beijing: Graduate School of Chinese Academy of Social Sciences*, pages 11–17.

Fang, Gao and Xu Shenghuan. 2000. Denominal verbs. *Foreign Language*, 2:7–14.

Fang, Le, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. 2019. Implicit deep latent variable models for text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3937–3947. https://doi.org/10.18653/v1/D19-1407

Fauconnier, Gilles. 1997. *Mappings in Thought and Language*. Cambridge University Press.

Fillmore, Charles. 1968. The case for case. *Universals in Linguistic Theory*, pages 1–89.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250. https://doi.org/10.1093/ijl/16.3.235

Franklin, Benjamin. 1789. *To Noah Webster, On New-Fangled Modes of Writing and Printing*.

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288. https://doi.org/10.1162/089120102760275983

Goldberg, Adele E. 2011. Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics*, 22(1):131–153. https://doi.org/10.1515/cogl.2011.006

Goldberg, Yoav and Jon Orwant. 2013. 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*, pages 241–247.

Graves, Alex and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Grice, Herbert P. 1975, Logic and conversation. In P. Cole and J. L. Morgan, editors, *Speech Acts*. Brill, pages 41–58. https://doi.org/10.1163/9789004368811_003

Hale, Ken and Samuel Jay Keyser. 1999. Bound features, merge, and transitivity alternations. *MIT Working Papers in Linguistics*, 35:49–72.

Hamilton, William L, 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. https://doi.org/10.18653/v1/P16-1141

Hermann, Karl Moritz, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1458. https://doi.org/10.3115/v1/P14-1136

Jespersen, Otto. 2013. *A Modern English Grammar on Historical Principles: Volume 7. Syntax*. Routledge. https://doi.org/10.4324/9780203715956

Kingma, Diederik P. and M. Welling. 2014. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Kingma, Durk P., Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Kingsbury, Paul R. and Martha Palmer. 2002. From TreeBank to PropBank. In *LREC*, pages 1989–1993.

Kisselew, Max, Laura Rimell, Alexis Palmer, and Sebastian Padó. 2016. Predicting the direction of derivation in English conversion. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 93–98. https://doi.org/10.18653/v1/W16-2015

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635.

Lakoff, George. 2008. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago press.

Lapata, Maria. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 63–70.

Lapata, Maria and Alex Lascarides. 2003. A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Li, Bai, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2020. Word class flexibility: A deep contextualized approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 983–994. https://doi.org/10.18653/v1/2020.emnlp-main.71

Lin, Kevin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial ranking for language generation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3158–3168.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mateu, Jaume. 2001. On the relational semantics of transitive denominal verbs. In I. Ortega-Santos, editor, *Current Issues in Generative Grammar*.

Minsky, Marvin. 1974. A framework for representing knowledge. In D. Metzing, editor, *Frame Conceptions and Text Understanding*. Berlin, Boston: De Gruyter, pages 1–25.

Nakov, Preslav and Marti Hearst. 2006. Using verbs to characterize noun-noun relations. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 233–244. https://doi.org/10.1007/11861461_25

Narayanaswamy, Siddharth, T. Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. 2017. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pages 5925–5935.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162

Press, Ofir, Amir Bar, Ben Bogin, Jonathan Berant, and Lior Wolf. 2017. Language generation with recurrent generative adversarial networks without pre-training. *arXiv preprint arXiv:1706.01399*.

Pustejovsky, James. 1991. The generative lexicon. *Computational Linguistics*, 17(4):409–441.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rigollet, Philippe. 2007. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(Jul):1369–1392.

Rumelhart, David E. 1975. Notes on a schema for stories. In D. G. Bobrow and A. Collins, editors, *Representation and Understanding*. Elsevier, pages 211–236. https://doi.org/10.1016/B978-0-12-108550-6.50013-6

Ruppenhofer, Josef and Laura Michaelis. 2014. *Linguistic Perspectives on Structure and Context: Studies in Honor of Knud Lambrecht*, chapter Frames and the interpretation of omitted arguments in English. https://doi.org/10.1075/pbns.244.04rup

Schank, Roger C. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3(4):552–631. https://doi.org/10.1016/0010-0285(72)90022-9

Semeniuta, Stanislau, Aliaksei Severyn, and Erhardt Barth. 2017. A hybrid convolutional variational autoencoder for text generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 627–637. https://doi.org/10.18653/v1/D17-1066

Shwartz, Vered and Ido Dagan. 2018. Paraphrase to explicate: Revealing implicit noun-compound relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1200–1211. https://doi.org/10.18653/v1/P18-1111

Si, Xianzhu. 1996. Comparative study of English and Chinese denominal verbs. *Foreign Language*, 3:54–58.

Subramanian, Sandeep, Sai Rajeswar, Francis Dutil, Chris Pal, and Aaron Courville. 2017. Adversarial generation of natural language. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 241–251. https://doi.org/10.18653/v1/W17-2629

Thompson, Cynthia A., Roger Levy, and Christopher D. Manning. 2003. A generative model for semantic role labeling. In *European Conference on Machine Learning*, pages 397–408. https://doi.org/10.1007/978-3-540-39857-8_36

Tribout, Delphine. 2012. Verbal stem space and verb to noun conversion in French. *Word Structure*, 5(1):109–128. https://doi.org/10.3366/word.2012.0022

Van de Cruys, Tim, Stergos Afantenos, and Philippe Muller. 2013. MELODI: A supervised distributional approach for free paraphrasing of noun compounds. In *Seventh International Workshop on Semantic Evaluation (SemEval 2013) in Second Joint Conference on Lexical and Computational Semantics (SEM 2013)*, pages 144–147.

van der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Vogel, Petra M. and Bernard Comrie. 2011. *Approaches to the Typology of Word Classes*. Walter de Gruyter.

Wang, Shan and Francis Bond. 2013. Building the Chinese open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

Xavier, Clarissa Castellã and Vera Lúcia Strube de Lima. 2014. Boosting open information extraction with noun-based relations. In *LREC*, pages 96–100.

Yu, Lei, Lana El Sanyoura, and Yang Xu. 2020. How nouns surface as verbs: Inference and generation in word class conversion. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1979–1985.