Chaining and historical adjective extension

Karan Grewal (karang@cs.toronto.edu)
Department of Computer Science, University of Toronto

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

Abstract

A hallmark of natural language is the innovative reuse of existing words. We examine how adjectives extend over time to describe nouns and form previously unattested adjective-noun pairings. Our approach is based on the idea of chaining that postulates word meaning to extend by linking novel referents to existing ones that are close in semantic space. We test this proposal by exploring a set of models that learn to infer adjective-noun pairings from historical text corpora for a period of 150 years. Our findings across three diverse sets of adjectives support a chaining mechanism that is sensitive to semantic neighbourhood density, best captured by an exemplar model of category extension. This work sheds light on the generative cognitive mechanisms of word usage extension.

Keywords: word usage extension; chaining; exemplar theory; generative model; adjectives

Introduction

Speakers of a language often need to describe new items driven by socio-cultural changes or technological innovations. One way of referring to a new item is to create a new word, but more often speakers choose to reuse an existing word (Ramiro, Srinivasan, Malt, & Xu, 2018). Here we explore how adjectives are reused and extended to pair with nouns over time and ask whether the processes of word usage extension can be understood in principled computational terms.

The topic of adjective-noun pairing has been traditionally tackled from the perspective of lexical composition. In particular, existing studies have explored what adjective-noun pairings are considered plausible (Lapata, McDonald, & Keller, 1999), and how adjectives can be combined with nouns sensibly either via probabilistic models (Lapata, 2001) or through ontological constraints (Schmidt, Kemp, & Tenenbaum, 2006). More recent work has also suggested that adjective-noun composition can be modelled using vector-space models such as Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). In these studies, adjectives are considered to be linear operators that act on nouns in a vector space that impose linear transformations (Baroni & Zamparelli, 2010; Boleda, Baroni, McNally, & Pham, 2013; Vecchi, Zamparelli, & Baroni, 2013; Vecchi, Marelli, Zamparelli, & Baroni, 2017) or conform to additive compositional models (Zanzotto, Korkontzelos, Fallucchi, & Manandhar, 2010). Despite this extensive line of work, sparse research has directly involved the dimension of time in the investigation of adjective-noun composition.

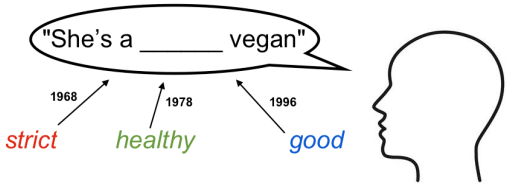

Figure 1: Example adjectives that emerged to describe *vegan* over the past half century.

Independent research in historical linguistics has suggested that adjective extension is non-arbitrary. For example, work on synaesthetic adjectives indicates that sensory terms such as those pertaining to sound, touch, and smell exhibit regular semantic change such that words from the same sensory domain tend to undergo parallel change in meaning (Williams, 1976). This line of inquiry takes an empirical approach and relies on historical dictionaries to characterize regularity in adjective usage, but to our knowledge there exists no formal computational treatment or large-scale evaluation of historical adjective extension.

Here we raise the question of how speakers choose to pair adjectives with nouns, particularly if such pairings have not yet appeared in the linguistic community. Figure 1 illustrates this problem. Given a noun such as *vegan*, different adjectives have been used as its modifiers over time. Although the historical order of adjective-noun pairings is influenced by non-linguistic or external factors, language users must still somehow come up with these novel adjective uses that are judged to be appropriate. We are interested in this question in its generality, which not only concerns a single noun like *vegan* but also any nouns, established or novel, in the lexicon.

We thus define the following problem: Given adjective-noun pairings at historical time t, can we predict novel adjective-noun pairings into the future $t+\Delta$? We characterize this problem in computational terms with the goal to understand the cognitive mechanisms that give rise to innovative word usage extension. Our basic premise is that the (temporal) choices of adjectives for a noun are not arbitrary, and as such, given knowledge of adjective uses in the past, one should be able to infer novel adjective-noun pairings into the future.

Our work is grounded in cognitive linguistic theories of chaining, which concern the cognitive mechanisms for category extension. Lakoff (1987) and other scholars (e.g., Malt, Sloman, Gennari, Shi, & Wang, 1999) have postulated that semantic categories grow via chaining, a process in which novel referents link to existing referents of a word due to proximity in semantic space. Chaining has been recently explored computationally in the historical extension of container names (Xu, Regier, & Malt, 2016), word sense extension (Ramiro et al., 2018), and more recently, the historical extension of numeral classifiers (Habibi, Kemp, & Xu, to appear). An important finding from these studies is that chaining as an extensional mechanism depends on semantic neighbourhood density, suggesting that historical meaning extension follows an incremental as opposed to abrupt process. In our study, we consider each adjective as a linguistic category and explore different mechanisms of chaining to predict how adjective usages grow to modify nouns that they have not previously co-occurred with.

Computational formulation

We formulate adjective extension as a temporal categorization problem. Given a noun n^*, information about its meaning at time t, a finite set of adjectives $\mathcal{A}$, and the historical adjective-noun pairings of adjectives $a \in \mathcal{A}$, we seek to predict which adjectives in $\mathcal{A}$ are most appropriate for n^* at time $t+\Delta$:

$$\begin{aligned} p\left(a|n^*\right)^{(t+\Delta)} &\propto p\left(n^*|a\right)^{(t)} p\left(a\right)^{(t)} \\ &= p\left(n^*|\{n\}_a^{(t)}\right) p\left(\{n\}_a^{(t)}\right), \end{aligned}$$

where we adopt the notation $\{n\}_a^{(t)}$ to reference the set of nouns that co-occur with adjective a at time t, i.e., the category extension of a. Note that when making predictions, we only consider *novel* adjective-noun pairs, i.e., adjectives $a \in \mathcal{A}$ that did not co-occur with n^* up to time t. Intuitively, this prediction task depends on two important sources of information: (i) what we know about other nouns that have co-occurred with a—captured in the likelihood term, and (ii) our belief about how dominant or common a is in the lexicon—captured in the prior. The likelihood and prior terms are based on information up to and including time t, but the posterior distribution of the left side of the equation describes what we wish to infer at time $t+\Delta$.

Likelihood function

As discussed earlier, semantic resemblance between words predicts that they are likely to be treated in similar ways and appear in the same adjective pairings (i.e., semantic chaining). Below we present models that operationalize chaining and semantic category extension in terms of different mechanisms (see Figure 2 for an illustration) when assigning to n^* adjectives that it is likely to co-occur with at future times. In particular, the chaining mechanisms are encapsulated in the likelihood term $p\left(n^*|a\right)^{(t)}$, which we formulate by drawing inspirations from work in machine learning (few-shot learning) and cognitive science (categorization theories).

Exemplar model According to exemplar theory (Ashby & Alfonso-Reese, 1995; Nosofsky, 1986), each noun $n \in \{n\}_a^{(t)}$ is an *exemplar* of adjective a. The degree of similarity between n^* and each exemplar noun n therefore determines the likelihood:

$$p\left(n^*|a\right)^{(t)} \propto \frac{1}{h\left|\{n\}_a^{(t)}\right|} \sum_{n \in \{n\}_a^{(t)}} \text{sim}\left(\vec{v}_{n^*}^{(t)}, \vec{v}_n^{(t)}\right),$$

where the *similarity* function sim measures how similar two nouns are and is defined as

$$\text{sim}\left(\vec{v}_{n^*}^{(t)}, \vec{v}_n^{(t)}\right) = \exp\left(-\frac{d\left(\vec{v}_{n^*}^{(t)}, \vec{v}_n^{(t)}\right)^2}{h}\right),$$

and $\vec{v}_{n^*}^{(t)}$ is a semantic representation of n^* at time t. In practice, $d\left(\cdot,\cdot\right)$ measures Euclidean distance between nouns and h is a kernel parameter that we learn. This model has been recently shown to predict the historical growth of Chinese numeral classifiers (Habibi et al., to appear), and here we examine if the same model might explain historical adjective extension. Note that this model is similar to performing kernel density estimation in semantic space defined by the likelihood function, and thus we use a kernel parameter h in the sim() function and also divide the resulting sum by h.

Prototype model Motivated by work in prototype theory (Rosch, 1975) with recent advancements in few-shot learning (Snell, Swersky, & Zemel, 2017), each adjective a has a *prototype representation* and n^*'s proximity to this prototype in semantic space determines how likely n^* is to co-occur with a at time $t+\Delta$. The likelihood is therefore $p\left(n^*|a\right)^{(t)} \propto \text{sim}\left(\vec{v}_{n^*}^{(t)}, \vec{p}_a^{(t)}\right)$ where the prototype $\vec{p}_a^{(t)}$ is computed as

$$\begin{aligned} \vec{p}_a^{(t)} &= \mathbb{E}\left[n \in \{n\}_a^{(t)}\right] \\ &\approx \frac{1}{\left|\{n\}_a^{(t)}\right|} \sum_{n \in \{n\}_a^{(t)}} \vec{v}_n^{(t)}. \end{aligned}$$

In essence, $\vec{p}_a^{(t)}$ is the centroid of all nouns vectors that co-occur with a. We also consider a variant of the prototype model in which the prototype representation for each adjective category remains static. That is, $\vec{p}_a^{(t)} = \vec{p}_a^{(t_0)}$ for all $t > t_0$ where t_0 is the base time (discussed later). We refer to this variant as the *progenitor model*.

***k*-nearest neighbors model** The basic idea is that examples within proximity of each other in semantic space exhibit similar properties and categories (Koch, 2015; Vinyals, Blundell, Lillicrap, Kavukcuoglu, & Wierstra, 2016). In a Bayesian framework, the k-nearest neighbors (k-NN) likelihood of n^*

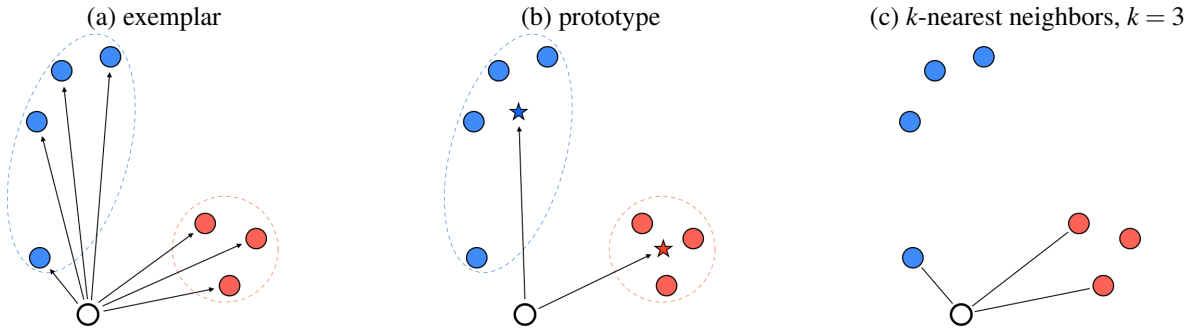

Figure 2: Illustration of the various chaining algorithms used to compute likelihood functions. The unshaded circle is the stimulus or the probe noun, red circles are nouns that have paired up with one particular adjective, and blue circles with another (although a noun may pair up with multiple adjectives). This example illustrates how the k-nearest neighbor model's prediction can differ from that of the prototype model based on the geometry of the semantic space.

pairing up with adjective a is proportional to whether its k closest neighbors $n_1, \ldots, n_k$ previously paired up with a, and inversely proportional to the size of category a. That is,

$$p(n^*|a)^{(t)} \propto \frac{1}{\left|\{n\}_a^{(t)}\right|} \sum_{j=1}^{k} \mathbb{1}\left[n_j \in \{n\}_a^{(t)}\right]$$

where the sum is over the k nouns closest to n^* in semantic space. When this likelihood is combined with the prior, the k-NN posterior probability amounts to n^*'s k closest neighbors voting (possibly more than once) for each of the adjectives that they previously paired up with. Note that this formulation of k-NN can be viewed as a "hard version" of the exemplar model where k is a discrete analog of the kernel parameter h. We report $k = 1$ and $k = 10$ in our experiments.

Prior distribution

We formulate a type-based prior $p(a)^{(t)}$ which gives how likely adjective a is to be paired with any noun based on its dominance in the lexicon, as discussed earlier. This formulation thus predicts that a's probability of appearing in a novel adjective-noun pairing is directly proportional to the number of unique nouns it has previously paired up with:

$$p(a)^{(t)} = \frac{\left|\{n\}_a^{(t)}\right|}{\sum_{a' \in \mathcal{A}} \left|\{n\}_{a'}^{(t)}\right|}.$$

This category-size-based prior serves as our baseline model when making adjective predictions for n^* at time $t + \Delta$, hence $p(a|n^*)^{(t+\Delta)} = p(a)^{(t)}$.

The rationale behind this choice of prior is as follows: if semantic chaining largely explains the emergence of novel adjective-noun pairs, then adjectives that have paired with more nouns have a higher *a priori* probability of "attracting" a given noun n^* via linking it to semantically similar nouns which are more likely to have previously co-occurred with a (Luo & Xu, 2018). This rich-get-richer process is also supported by work on how semantic networks grow through preferential attachment (Steyvers & Tenenbaum, 2005). Furthermore, this prior can be integrated with the likelihood functions specified in a full Bayesian model.

Semantic space

The chaining algorithms described above operate in semantic space. We used Word2Vec-based representations commonly used in natural language processing for distributed semantics (Mikolov et al., 2013). Note that word co-occurrence distributions are constantly changing and therefore the semantic space needs to be updated to capture information only up to time t. For this reason, we use diachronic (historical) Word2Vec embeddings (Hamilton, Leskovec, & Jurafsky, 2016) where at each time t, the embedding for each noun is based solely on its co-occurrence statistics at time t, and all past and future co-occurrences are ignored. Hence, the predictions made by all models are in a sense "zero-shot", or without access to semantic space in the future.

Historical data of adjective-noun uses

We extracted a large database of historical adjective-noun uses over the past 150 years (1850 - 2000). We collected these data from the Google Books corpus (Lin, Michel, Aiden, Brockma, & Petrov, 2012) which contains transcriptions of books written between 1800 and 2000. Within Google Books, the English All (ENGALL) corpus accounts for 8.5×10^{11} tokens and roughly 4% of all books ever published. The size of the ENGALL corpus is likely to reflect how the English language has changed over the past centuries, and moreover making our adjective-noun co-occurrence dataset suitable for evaluating hypotheses about word usage extension.

We collected adjective-noun co-occurrence counts from the ENGALL corpus. First, we extracted all bigrams from the ENGALL corpus in which the first token is an adjective and the second is a noun (by specifying POS tags) along with the corresponding timestamp. As the corpus is likely to contain noise, we standardized the set of nouns and adjectives by only considering ones contained in WordNet (Miller, 1995), which gives approximately 67k nouns and 14k adjectives.

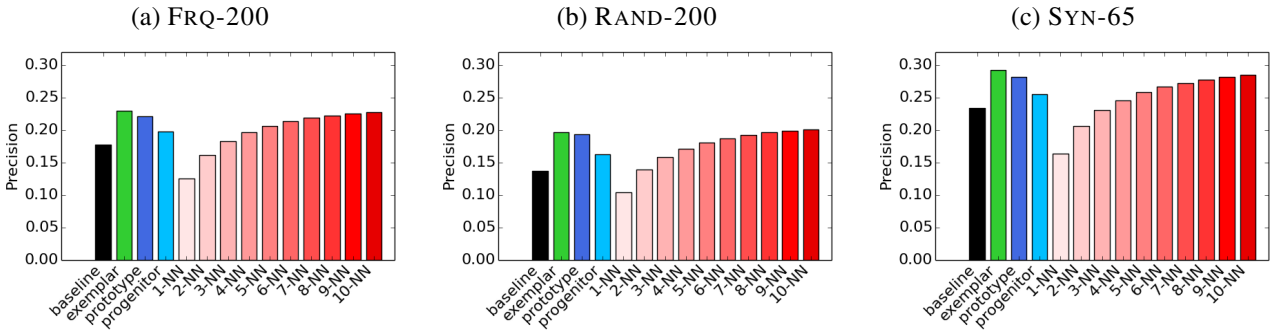

Figure 3: Aggregate precision accuracy for all models (including k-NN from $k = 1$ to $k = 10$) across all time periods on each of our three adjectives sets.

We collapsed raw co-occurrence counts into decadal bins by choosing $\Delta = 10$ years. This yielded our adjective-noun pairings dataset which consists of entries of the form (a, n, count, t). In each decade t, we used a pre-trained Word2Vec language model as the semantic space. For our analyses we worked with a subset of the collected data (discussed in the next section), due to both considerations of sampling and computational feasibility. To account for semantic change across decades, we used diachronic Word2Vec embeddings which were also trained using the ENGALL corpus. Hamilton et al. (2016) also chose to construct diachronic Word2Vec embeddings decade-by-decade for similar reasons.

We now describe three adjective sets $\mathcal{A}$. The purpose of testing our models on three different adjective sets is to obtain representative samples of adjectives, and to ensure our hypotheses are robust to choice of groups of adjectives.

Frequent adjectives. We use multiple ways to construct $\mathcal{A}$ such that it covers a broad scope and we show our results are reproducible and agnostic to choice of adjectives. To construct a set of 200 adjectives that cover a broad range of descriptions, we first collected word vectors of all adjectives in the Google Books corpus using a pre-trained Word2Vec model. Next, we clustered the adjectives into 20 clusters and picked 10 adjectives from each to construct our set $\mathcal{A}$ of 200 adjectives. Adjectives were sampled from each cluster based on their lexical frequency, and only competed against other adjectives within the same cluster during sampling. We refer to this set as FRQ-200, with examples shown in table 1.

Random adjectives. To ensure that the sampling scheme for choosing $\mathcal{A}$ is not biased towards token frequencies, we also constructed another set of 200 adjectives by repeating the clustering step as described above, but replaced frequency sampling with uniform sampling. We refer to this dataset as RAND-200. As Table 1 shows, adjectives drawn from the same cluster are semantically similar between FRQ-200 and RAND-200, but less common in the latter set.

Synaesthetic adjectives. In addition to choosing common adjectives, we also consider the set of *synaesthetic adjectives* (SYN-65) defined by Williams (1976), as a more focused domain. This set includes 65 adjectives[1] that exhibit regularity in their extension patterns. For instance, Williams shows how adjectives that originally described touch perceptions have since extended to describe color (e.g., *warm cup* $\longrightarrow$ *warm color*), and adjectives that originally described color started to describe sound (e.g., *clear blue* $\longrightarrow$ *clear voice*). We will refer to this set as SYN-65.

All data and code from our analyses are publicly available[2].

Table 1: A comparison of some adjectives in FRQ-200 and RAND-200 grouped according to the cluster they were drawn from. Notice that the clusters align semantically, however the adjectives in FRQ-200 are more frequently represented in the English lexicon than those in RAND-200.

FRQ-200	RAND-200	FRQ-200	RAND-200
Asian	*Hungarian*	*polite*	*chatty*
Christian	*Thai*	*intelligent*	*unorthodox*
American	*Cornish*	*passionate*	*amiable*
European	*Catalan*	*energetic*	*communicative*

Results

We tested our models on their ability to predict which adjectives $a \in \mathcal{A}$ would pair up with a given noun n^* in decade $t+\Delta$ given all information about n^* up to and including decade $t > t_0$, where t_0 is the base decade. This information includes co-occurrences between all nouns n and adjectives $a \in \mathcal{A}$ at or before decade t as well as time-dependent word embeddings at each decade, taken from Hamilton et al. (2016). We chose t_0 =1840s, yet to build an initial lexicon, our dataset of adjective-noun co-occurrences dates back to the 1800s. The 1860s was the first decade for which we report predictions,

[1]There are in fact 64 unique adjectives in this set and WordNet captures only 61 of these. See Williams (1976) for a comprehensive list.

[2]Code and data are available at `https://github.com/karangrewal/adjective-extension`.

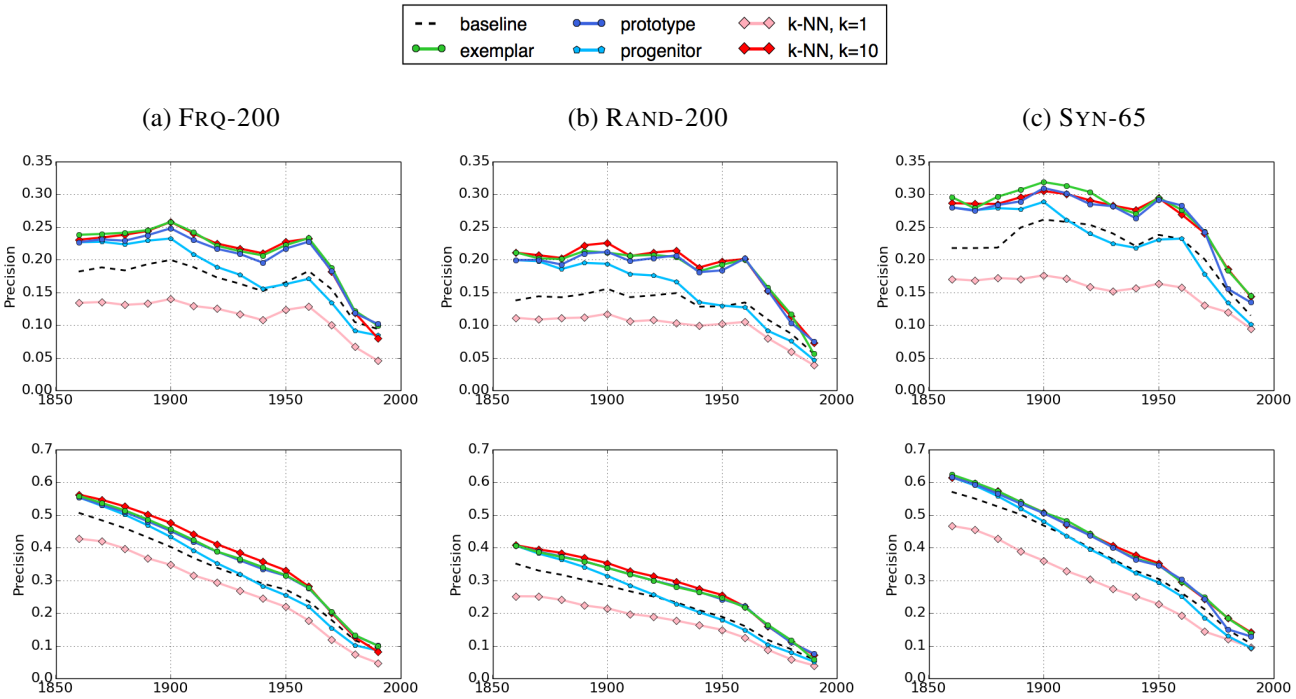

Figure 4: Model predictive accuracy on the FRQ-200, RAND-200, and SYN-65 adjective sets. **Top row**: Predictive accuracy when only novel adjective-noun pairs in the following decade are considered. **Bottom row**: Predictive accuracy when all future adjective extensions are considered.

as we used the 1850s as our "training decade" to estimate the kernel parameters for the exemplar and prototype models.

We define the co-occurrence (a, n^*) to be novel in decade $t+\Delta$ if and only if (i) a co-occurred with n^* in decade $t+\Delta$ beyond a certain threshold (which we set to 2), and (ii) a never co-occurred with n^* beyond that threshold in any decade $t' \leq t$. Using these criteria allowed us to eliminate noise from co-occurrence statistics. Given a noun n^*, each model's output was a categorical distribution $p(a|n^*)^{(t+\Delta)}$ over all adjectives $a \in \mathcal{A}$. The model was then scored on its precision accuracy on the set of adjectives that first co-occurred with n^* in decade $t+\Delta$. That is, if n^* co-occurred with m new adjectives in $\mathcal{A}$ in decade $t+\Delta$, then we took the top m adjectives with the highest posterior probability that previously didn't co-occur with n^* as the set of retrieved positives. We report total precision accuracy for all models and also use this as an objective to learn all kernel parameters. Also, we considered two types of predictive tasks when making predictions for noun n^* in decade t: taking as ground truth adjectives that co-occur with n^* (1) specifically in decade $t+\Delta$, and (2) any future decade $t' > t$ up to the terminal decade 1990s.

Next, we discuss results from our experiments for differently sampled adjective sets $\mathcal{A}$. As Figure 3 shows, the exemplar model has the highest predictive performance, followed closely by the 10-NN and prototype models. The exemplar, prototype, and 10-NN models are perform substantially better than the baseline. We hypothesized that the 10-NN model would be not better than the exemplar model as the kernel parameter is a continuous analog of k and is optimized for precision, and this is indeed the case. The progenitor model, a variant of the prototype model with "static" prototypes determined in decade t_0, becomes considerably worse than the prototype model with time. This relationship between the prototype and progenitor models that we observe indicates that if the prototype model is the closest underpinning of adjective extension, then $\{n\}_a^{(t)}$ largely influences which nouns adjective a will extend to and that each adjective category "center" updates once novel adjective-noun pairings are formed.

Further results with year-over-year accuracy are shown in Figure 4. The predictive accuracy falls in later decades since there are fewer novel adjective-noun pairings to predict. Examples of predictions made by our models are provided in Table 2. Our results hold generally across all 3 adjective sets, and they suggest that semantic neighborhood density is an important factor contributing towards adjective extension as the exemplar and 10-NN models achieve the overall best predictive accuracy.

Discussion and conclusion

We have presented a computational study of historical adjective extension, examined through a large dataset of adjective-noun pairings over the last 150 years. We have focused on exploring different mechanisms of semantic chaining to predict adjective-noun pairings over time. Our results indicate that among the different model variants, the exemplar model

Table 2: Examples of model predictions on the FRQ-200 adjective set. Adjectives in bold font indicate true positives retrieved by models. We present predictions for nouns *cigarette*, *alcohol*, and *Vietnam* as the adjectives they first pair with in the 1880s, 1920s, and 1960s respectively reflect sentiment (e.g., *social cigarette*) or historic events (e.g., *illegal alcohol* due to prohibition, *American Vietnam* due to the Vietnam war).

noun & decade	*cigarette*, 1880s
new adjectives	*better*, *modern*, *several*, *excessive*, *American*, *social*
baseline prediction	*original*, *particular*, *English*, *natural*, *perfect*, ***modern*** (1/6)
exemplar prediction	*black*, *red*, *English*, *poor*, *original*, *particular* (0/6)
prototype prediction	*red*, *black*, *dry*, *warm*, *cold*, *English* (0/6)
10-NN prediction	*original*, *warm*, *particular*, *red*, *English*, *dry* (0/6)
noun & decade	*alcohol*, 1920s
new adjectives	*female*, *analogous*, *red*, *bitter*, *marked*, *illegal*
baseline prediction	*perfect*, *extraordinary*, *moral*, *physical*, *western*, *christian* (0/6)
exemplar prediction	***red***, *moral*, *artificial*, *dense*, *perfect*, ***marked*** (2/6)
prototype prediction	*artificial*, *perfect*, ***marked***, ***red***, *physical*, *moral* (2/6)
10-NN prediction	***red***, *moral*, *dense*, *perfect*, ***analogous***, *artificial* (2/6)
noun & decade	*Vietnam*, 1960s
new adjectives	*western*, *tropical*, *eastern*, *colonial*, *particular*, *more*, *top*, *poor*, *American*
baseline prediction	*same*, ***more***, *great*, ***particular***, ***American***, *different*, *natural*, *human*, *English* (3/9)
exemplar prediction	***western***, ***eastern***, ***more***, ***particular***, *great*, ***colonial***, *inner*, *same*, ***poor*** (6/9)
prototype prediction	*great*, *same*, ***western***, ***more***, ***American***, ***eastern***, ***particular***, *European*, *French* (5/9)
10-NN prediction	***western***, ***eastern***, ***more***, ***tropical***, ***colonial***, *great*, *better*, *inner*, ***particular*** (6/9)

tends to perform the best in predicting the historical data, followed closely by related models including 10-NN and prototype models. These findings support our overall hypothesis that semantic neighborhood density influences how novel adjective-noun pairings emerge, although the distinction between the exemplar model and the competitive models (e.g., prototype model) appears to be quite minimal for drawing any strong conclusion from this initial investigation. Nevertheless, all the models we examined perform considerably better than the baseline model that extends adjectives by a majority-vote mechanism, and this finding is consistent through the historical period of investigation. Our work is thus consistent with existing work on chaining on its role as a key mechanism in the growth of linguistic categories (Lakoff, 1987; Malt et al., 1999; Xu et al., 2016; Ramiro et al., 2018), and we extend these studies to explaining the usage extension of adjectives.

Our investigation has its limitations. First, our operationalization of chaining depends crucially on semantic similarity. One drawback of this assumption is that although chaining mechanisms may retrieve nouns that are similar to a probe noun, there is no independent mechanism of checking whether the adjective-noun pairing is plausible. That is, our implementation of chaining does not explicitly "perform a check" as to whether a predicted adjective-noun pairing is sensible. This perhaps explains partly why our models make predictions such as *moral alcohol* (see Table 2) which is nonsensical with respect to any known sense of the adjective *moral*, and such a pairing in fact has never been attested. As adjectives accumulate novel senses and uses, the set of possible nouns they can pair with will vary due to external factors additionally to chaining. Here we acknowledge this limitation, but also conjecture any potential method that can discern nonsensical adjective-noun compositions should in principle yield better performance in the predictive models.

Second, we have assumed that the distributed semantic representations are adequate to capture the meaning of nouns. In particular, we used Word2Vec to capture distributional meaning of words from linguistic context or usage, but many other variants of distributed semantic models are available. More importantly, perceptual (e.g., visual) features might be especially important for nouns that are concrete and imageable, and our current construction of the semantic space might not capture these features. There is some empirical evidence to suggest that adjective usage prediction might benefit from visual information. For instance, Lazaridou, Dinu, Liska, and Baroni (2015) proposed cross-modal mappings between visual and linguistic representations that assign adjective labels to visual inputs, and Nagarajan and Grauman (2018) followed up by learning a linear mapping that predicts adjective descriptors based on a visual input. However, one limiting factor of these cross-modal approaches is that they may not be relevant to predicting adjective pairings with abstract nouns where perceptual grounding is difficult to establish.

To conclude, our work provides a starting point for exploring the composition of adjectives and nouns through the lens of historical language change and probabilistic algorithms. Our approach provides important clues to the generative cognitive mechanisms that may underlie word usage extension, and should stimulate future work on the interaction of internal and external factors in shaping innovative language use.

Acknowledgments

We thank Barend Beekhuizen, Charles Kemp, and Sammy Floyd for helpful discussion. We also thank Amir Ahmad Habibi for sharing data and code. KG is supported by a Bell Graduate Scholarship and a Vector Scholarship in Artificial Intelligence. YX is funded through an NSERC Discovery Grant, a SSHRC Insight Grant, and a Connaught New Researcher Award.

References

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

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing.*

Boleda, G., Baroni, M., McNally, L., & Pham, N. (2013). Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 7th international conference on computational semantics.*

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

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th annual meeting of the association for computational linguistics.*

Koch, G. (2015). *Siamese neural networks for one-shot image recognition*. Unpublished master's thesis, University of Toronto.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.

Lapata, M. (2001). A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the 2nd meeting of the north american chapter of the association for computational linguistics* (pp. 63–70).

Lapata, M., McDonald, S., & Keller, F. (1999). Determinants of adjective-noun plausibility. In *Proceedings of the 9th conference of the european chapter of the association for computational linguistics.*

Lazaridou, A., Dinu, G., Liska, A., & Baroni, M. (2015). From visual attributes to adjectives through decompositional distributional semantics. *Transactions of the Association for Computational Linguistics*, *3*.

Lin, Y., Michel, J.-B., Aiden, E. L., Brockma, J. O. W., & Petrov, S. (2012). Syntactic annotations for the google books n-gram corpus. In *Proceedings of the 50th annual meeting of the association for computational linguistics.*

Luo, Y., & Xu, Y. (2018). Stability in the temporal dynamics of word meanings. In *Proceedings of the 40th annual conference of the cognitive science society.*

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

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

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *39*(11), 39–41.

Nagarajan, T., & Grauman, K. (2018). Attributes as operators: Factorizing unseen attribute-object compositions. In *Proceedings of the 15th european conference on computer vision.*

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, *115*, 2323–2328.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*, 192–233.

Schmidt, L. A., Kemp, C., & Tenenbaum, J. B. (2006). Nonsense and sensibility: Inferring unseen possibilities. In *Proceedings of the 28th annual conference of the cognitive science society.*

Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical networks for few-shot learning. In *Advances in neural information processing systems.*

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41–78.

Vecchi, E. M., Marelli, M., Zamparelli, R., & Baroni, M. (2017). Spicy adjectives and nominal donkeys: Capturingsemantic deviance using compositionality in distributional spaces. *Cognitive Science*, *41*(2), 102–136.

Vecchi, E. M., Zamparelli, R., & Baroni, M. (2013). Studying the recursive behaviour of adjectival modification with compositional distributional semantics. In *Proceedings of the 2013 conference on empirical methods in natural language processing.*

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems.*

Williams, J. M. (1976). Synaesthetic adjectives: A possible law of aemantic change. *Language*, *32*, 461–78.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*, 2081–2094.

Zanzotto, F. M., Korkontzelos, I., Fallucchi, F., & Manandhar, S. (2010). Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd international conference on computational linguistics.*