Chaining Algorithms and Adjective Extension

by

Karan Raj Singh Grewal

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Abstract

Chaining Algorithms and Adjective Extension

Karan Raj Singh Grewal

Master of Science

Graduate Department of Computer Science

University of Toronto

2020

A hallmark of natural language is the innovative reuse of existing words. In this thesis, we take a computational perspective to examine how adjectives extend over time to describe nouns and form previously unattested adjective-noun pairs. We hypothesize that the underlying mechanisms that govern how adjective-noun pairs emerge exhibit regularity, and that this phenomenon is not entirely random. 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 probabilistic models that learn to infer novel adjective-noun pairs from historical text corpora that span a period of 150 years. Our findings across three diverse sets of adjectives support a chaining mechanism sensitive to local semantic neighborhoods, and this finding aligns with what researchers studying language change in other domains have found. Our work sheds light on the generative cognitive mechanisms that may underlie word usage extension.

Acknowledgements

There are various individuals who have made this thesis possible, and I would like to acknowledge all of them. To start, I would like to thank the following faculty members to whom I am grateful. My advisor Yang Xu introduced me to cognitive modelling and invested a large portion of his time in my progress and development. Tom Griffiths gave me invaluable opportunities that have largely shaped me as a researcher, and also devoted many hours towards my progress. Barend Beekhuizen contributed to my thesis work by reading early drafts and providing helpful feedback, and I thank him for his service as the second reader of my thesis.

Next, I have been fortunate to share my graduate school experience with Renato Ferreira and Zhewei Sun as we all started together, encountered similar challenges, and celebrated our successes together. During a large portion of this time, it was an absolute pleasure to work with Bill Thompson and Josh Peterson who were mentors to me and from whom I learned tremendous amounts. I would also like to mention and give thanks to the following individuals whom I enjoyed collaborating with: Amir Ahmad Habibi, Robert Hawkins, and Sammy Floyd.

Finally, I want to thank my parents for their unconditional support, my two older siblings who also provided me with parental wisdom, and my friends for being by my side.

Contents

Chapter 1

Introduction

Natural language expresses a potentially infinite set of ideas with a finite lexicon. Speakers of a language often encounter cultural evolution, technological innovations, and other factors, and these changes ultimately drive languages to adapt under limited cognitive resources [18, 21].

However, languages do not necessarily need to create a new term each time they express a novel concept or idea, since the size of the vocabulary would grow rapidly, thus placing a strain on learning a very large lexicon. Instead, one alternative solution to expressing novel concepts is the creative reuse of existing words. The most common type of word reuse is perhaps the accumulation of word senses over time, such as how *face* originally referred to the body part, subsequently extending to surfaces of inanimate objects, and eventually the action of taking on challenges. In this thesis, we explore word usage extension and how words appear in novel contexts. One particular domain of word reuse is *adjective extension*: the process by which adjectives pair with nouns that they previously did not encounter to form novel adjective-noun combinations. Adjectives serve the primary function of expressing the attributes of a wide range of nouns. Indeed, one of the most common uses of adjectives in English is the modification of nouns, e.g., "strict$_{\text{ADJ}}$ person$_{\text{NOUN}}$". A natural problem faced by speakers is how to pair any adjective with nouns that it previously did not encounter. For instance, although we may most often use strict to refer to the behavior of a person, we can also use it to describe the variance in a diet (*strict diet*) or the necessity of a criterion (*strict criterion*) as the adjective *strict* has extended over time and developed new senses in order to be applicable to these other nouns since its original use.

An alternative way of looking at adjective extension is to ask which adjectives will become modifiers for a given noun over time, and Figure 1.1 illustrates this problem. Instead of fixing an adjective and observing the nouns it will extend to, we fix a noun and are interested in which adjectives that

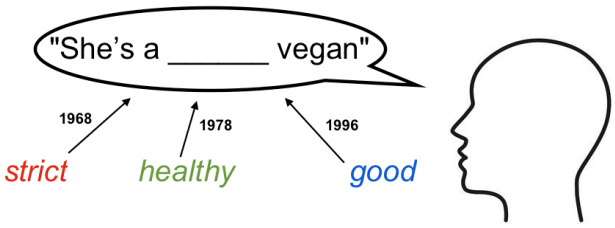


Figure 1.1: The emergence of adjectives that co-occur with *vegan* over the past half century.

previously did not pair with said noun are likely candidates to form a novel adjective-noun combination. One interesting example that we use as a case study is a subset of the adjectives that have modified *vegan* (i.e., "she's a *vegan*") during the last half-century, especially since veganism has been a controversial subject. Animal rights, ethics, and environmental sustainability all drew little attention from the public eye during the most part of the twentieth century in comparison to modern day due to lack of public interest in these topics. However, they gradually gained importance throughout the 1970s, 1980s, and 1990s, and as veganism is central to these topics, the choice of adjectives that English speakers employed to pair with *vegan* perhaps reflects this cultural shift. Prior to the 1970s, we observe the novel pairings *strict vegan* which has carries a low degree of positive sentiment, but during the period in which we suspect a change in popular views, we observe the novel pairs *healthy vegan* and *good vegan* emerge, and both these adjectives carry much greater positive sentiment towards veganism[1]. This example highlights how (a) adjective-noun pairs can emerge at different times, and (b) social factors influence which adjectives are likely to pair with nouns such as *vegan*.

Given these observations, we naturally ask whether the novel contexts in which adjectives and nouns co-occur over time exhibit any regularity or patterns. In doing so, we also wish to make two clarifying remarks about the nature of our problem. First, we emphasize that adjective extension in our study is not equivalent to adjectives accumulating novel senses as modifiers, but instead we focus on which adjectives and nouns co-occur together over time. For instance, *healthy* and *good* both extended to *vegan*, but the way in which both adjectives are used as modifiers is not novel, instead the context is, as it gives rise to previously unseen adjective-noun pairings. Predicting the emergence of novel adjective senses is an entirely different problem that we do not focus on here, however it may benefit from our approach since new adjective senses generally imply new contexts as well. Second, while it remains true that external influences such as technological innovations and changes in social perspectives are possibly unpredictable and play a large role in the formation of novel adjective-noun pairs, we argue that they

[1]Source: `https://books.google.com/ngrams`.

exert a domino effect on adjective usage. As such, we posit that the emergence of the combination *healthy vegan* followed by *good vegan* is not a coincidence, but rather frequent usage of the former by English speakers led to the latter. In a sense, we aim to show that external factors may guide the change of the formation of novel adjective-noun pairs, but the cognitive mechanisms that explain most novel pairs has an underpinning, and that is what we wish to model.

Our basic premise is that the temporal choices of adjectives for a noun are not arbitrary, and given knowledge of adjective uses in the past, one might be able to predict novel adjective-noun pairs into the future. In particular, we explore the idea of adjective extension as chaining, a computational mechanism that we describe in Chapter 2, and hypothesize it can capture the underlying process that generates new adjective-noun pairings. This idea has already been generalized to other domains of word usage extension, namely word sense extension [45], the evolution of container names [66], and more recently, the historical extension of numeral classifiers [14]. We provide a computational model of adjective extension through chaining in semantic space based primarily on proposals of semantic chaining [20, 29, 45, 66]. In particular, we test exemplar, prototype, and nearest neighbors models that learn to infer novel adjective-noun pairs over time and show that chaining is a core component as these models perform much better than a simple category-size-based prior. We aim to uncover the generative cognitive mechanisms that underlie word usage extension.

This thesis is structured as follows: in Chapter 2, we discuss past work in psychological, linguistic, and computational domains that are requisite to modelling adjective extension, such as chaining. In Chapter 3, we formalize the problem of adjective extension, present a probabilistic model that learns to infer novel adjective-noun pairs, and go into detail about computational accounts of semantic chaining. In Chapter 4, we present a historical dataset of adjective uses that spans 150 years and which we use as an empirical test bed for our predictive models. Finally, in Chapter 6, we reflect on our results, modelling decisions, and future avenues that may address the shortcomings of our approach.

Chapter 2

Previous Work

In this chapter, we discuss computational work of linguists, psychologists, and cognitive scientists that this thesis builds on. We discuss the growth of linguistic categories and chaining, as word usage extension relies heavily on these principles. We also delve into how computational models have assessed the plausibility of adjective-noun pairs, and finally we touch on some hypotheses regarding rules that govern how adjectives extend.

2.1 Linguistic growth and chaining

As speakers of natural language, we possess a finite vocabulary and thus have a finite number of adjectives that we can choose to pair with any given noun. We can therefore interpret adjective extension as categorizing nouns into one of many adjective categories. For instance, we may categorize a *vegan* person as being *healthy* which naturally permits the phrase *healthy vegan* to become more regular. When thinking about adjectives as categories, however, nouns can be assigned to (or classified amongst) one or more adjectives, and these adjective categories naturally grow (or possibly shrink) over time as novel adjective-noun pairs emerge.

In fact, categorization applies to much of language change. Linguists such as George Lakoff have proposed that linguistic categories (such as the set of nouns that an adjective frequently pairs with) grow over time through *semantic chaining*, a process in which categories attract new stimuli based on semantic similarity [7, 12, 20, 29]. This can be thought of as a set of category members that "reach out" to nearby stimuli in semantic space and grow the category by recruiting one or more nearby stimuli. Lakoff used the term chaining to describe this growth process as it forms chains that extend outward from the initial set of category members. This theory has been applied to modelling the extension of

word senses [45], container names [66], slang [58], and Chinese numeral classifiers [14] while performing considerably better than chance which suggests that semantic chaining plays a critical role in growing linguistic categories.

The way in which the chaining mechanism grows linguistic categories can vary. In Lakoff's work, he introduces the idea of centrality, i.e., that each category has a "center". Later work from other scholars furthers the argument for a center and that linguistic categories grow by radiating outwards from their respective centers [26], and even suggests a linguistic category may have multiple such centers of importance [40]. The most notable work along this direction is Eleanor Rosch's prototype theory [48] in which she refers to these category centers as *prototypes*. Rosch argues that the semantic similarity between category prototypes and novel stimuli directly influences which category novel stimuli are assigned to. In later work, she suggests that category prototypes need not be semantic representations for the categories [49], yet most work on prototype theory makes representational claims.

The prototype-based growth paradigm assumes that categories grow by radiating outwards from their prototypes, forming a web-like pattern in semantic space. However, this is not the only type of category growth; chaining can take several forms. Consider, for instance, a growth mechanism in which one member of the category is chosen at random and its semantic similarity with novel stimuli determines how those stimuli are categorized. A prototype-based growth process can be imagined as category members radiating outwards from the specific category prototype around which all stimuli cluster, whereas this specific growth mechanism would likely be illustrated by trees branching outwards at almost every node, and thus forming long chains in a somewhat-unorganized manner in semantic space.

An alternative chaining mechanism that informs how linguistic categories can grow is based on exemplar theory. Rather than assuming a category prototype or center controls the category's growth trajectory, the exemplar model takes all category members, termed *exemplars*, into account when categorizing novel stimuli. The key difference here is that semantic neighborhood, or semantic neighborhood density, matters; multiple configurations of category exemplars can yield the same category prototype and ultimately result in the same categorization decision. However, this is not true of exemplar-based growth since each exemplar's semantic similarity to a novel stimulus influences the categorization decision. Robert Nosofsky's Generalized Context Model (GCM) [37] formalizes exemplar theory and is perhaps the most well-known version of an exemplar model of categorization. The GCM has generally outperformed all other psychological models of categorization, including Rosch's prototype model, and exemplar accounts of word modelling have been also applied to several aspects of language including phonetics, phonology, morphology, word senses, and constructions [6, 17, 44, 46, 53]. Figure 2.1 illustrates

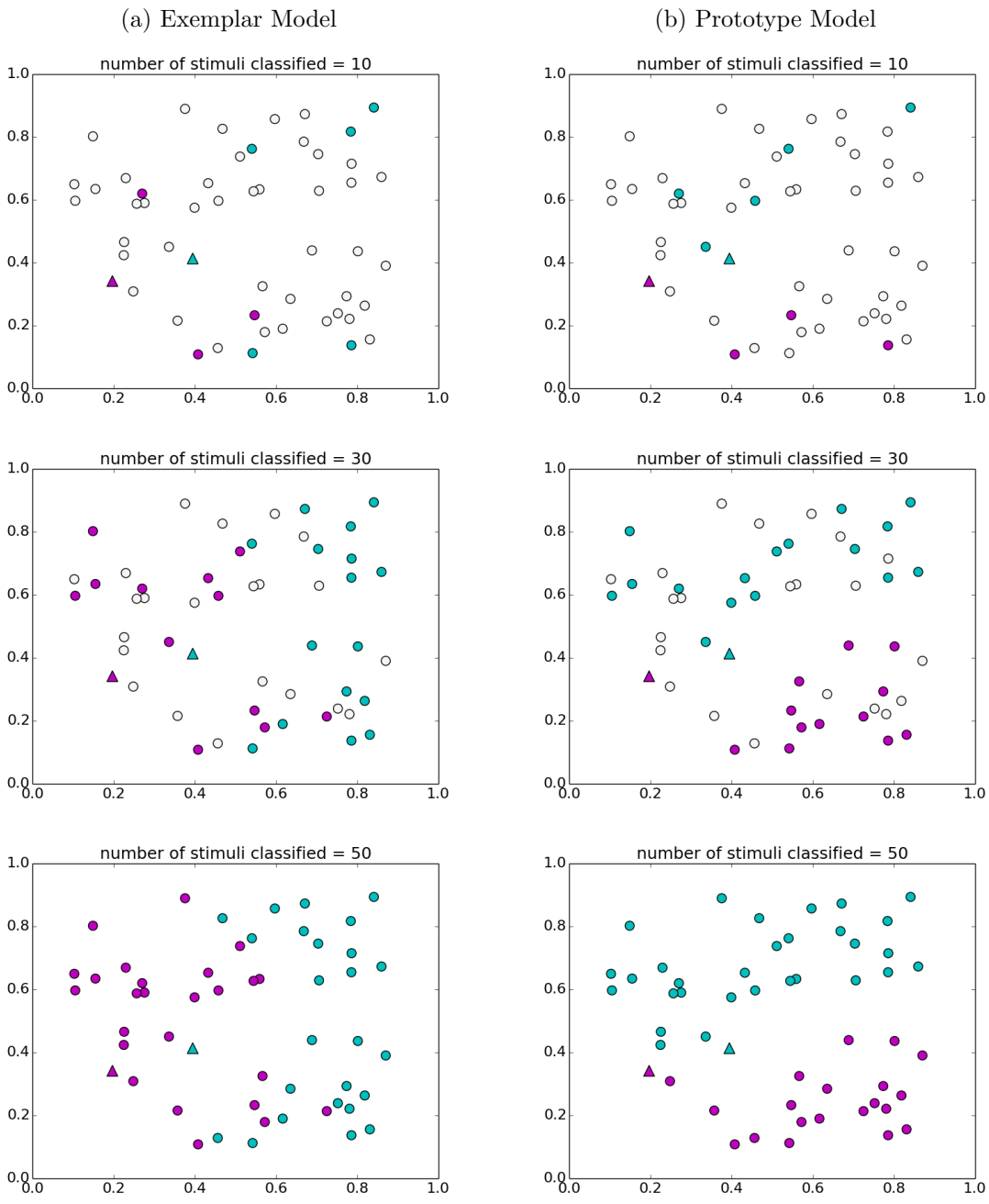


Figure 2.1: The resulting shape of categories when growing using exemplar (left) versus prototype (right) models. The triangle represents the first exemplar in each category as both categories grew from just one member. All the novel stimuli (circles) are initially ordered randomly and appear in a sequence (independent of proximity to either category) for classification. In each of the plots, the grey circles have not yet been classified whereas all the colored circles have been classified. This simulation shows how category assignments can differ based on the category growth model.

how the growth of two categories can differ when using the exemplar versus prototype model.

Beyond prototype and exemplar theories, nearest neighbor chaining has also proven effective at

growing linguistic categories. A nearest neighbor approach takes all exemplars of a category and picks out the one that is closest to a novel stimulus. The semantic similarity between the nearest exemplar and novel stimulus then determines the growth of a category. This method essentially builds a minimal spanning tree in semantic space and researchers have shown that this type of chaining best explains how word senses emerge over time [45]. Similarly, two earlier studies also found nearest neighbor chaining to best summarize how container names evolve across languages [54, 66].

2.2 Adjective-noun composition

As we're primarily interested in adjective-noun pairs and their temporal co-occurrences, it's important to first consider their acceptability. Not all adjectives in the English lexicon can pair with any noun, and this is best illustrated by Noam Chomsky's famous nonsensical phrase *colorless green ideas sleep furiously* [8]. For this reason, many researchers have taken a computational approach to investigate what makes adjective-noun pairs sensible. Note that there's a stark difference between *sensible* and *attested*: for instance, *homosexual vegan* is perhaps unattested and it's likely that most English speakers have never come across this phrase, but it's easy to imagine. On the other hand, *slippery vegan* is also unattested, but what does it even mean for a *vegan* to be *slippery*? This is a nonsensical composition.

In this section, we focus on computational approaches to adjective-noun composition. As speakers are better than any computational model at determining sensible composition, a simple way to assess plausibility is by studying human judgments. Interestingly, human judgments correlate strongest with the lexical frequency of adjective-noun pairs among various corpus-based variables, suggesting lexical frequency may be the driver of plausibility (or vice-versa) [23]. Similarly, when adjectives are polysemous or context-sensitive, a probabilistic model can identify the sense of the adjective and this also reflects what human subjects would intuit its sense to be [22]. Apart from assessing composition based on human judgments, does there exist a formal mathematical model that can distinguish acceptable adjective-noun pairs from nonsensical ones? One group of researchers assumed properties of noun objects are strictly hierarchical and they developed a Bayesian framework that learns to infer sensible compositions while adhering to ontological constraints [51]. For instance, since *vegans* can be *homosexual* but not *slippery*, and *ice* can be *slippery* but not *homosexual*, this implies no noun object can possibly be both *homosexual* and *slippery*.

Vector space models of semantics, such as Word2Vec [31], can also provide insight into composition. One simple method is to measure how similar a query noun is to the prototype for an adjective category as an assessment of adjective-noun fit. Although this has not explicitly been tried, an analogous approach

was used to determine how likely noun objects are to perform certain actions, where verbs are treated as categories and the nouns paired with a given verb are used to estimate the prototype [9, 10, 39]. This approach is almost identical to our prototype model, however one major difference is that we explore the temporal aspect of adjective-noun pairs rather than holding the time variable fixed. Another trend popularized by Marco Baroni has been to treat adjectives as operators that apply to noun vectors and result in a transformation that should subsequently yield the distributed representation of the adjective-noun phrase. This general line of work self-identifies as "compositional distributional semantics" since adjectives and nouns are are represented in vector spaces based on their distributional meaning and composed in various ways. For instance, by treating adjective-noun pairs as single tokens and estimating their distributional representations within a large corpus of text, least squares regression is best able to reconstruct the joint context using the individual adjective and noun tokens as compared with other compositions such as addition, pointwise multiplication, etc. between adjective and noun representations [13]. The distribution of an adjective-noun phrase can also be successfully modelled through various linear methods to compose representations of adjectives and nouns; for the most part, they treat adjectives as linear operators on nouns in a vector space [3, 5, 59], or more specifically as additive compositional models [67]. Pushing composition even further, one study showed how to model the distribution of adjective-adjective-noun phrases in a sensible way [60]. Researchers have also evaluated how various compositions align with human judgments [33]. In general, all these aforementioned approaches use matrix-vector composition and can be leveraged by neural network models to learn more sensible parses of phrases [56].

One possible argument against these vector space approaches based on distributional semantics is that humans do not only build mental models of the world regarding which adjective-noun compositions are sensible from large corpora of text, but also from the visual world. Recent studies have explored the composition of nouns and adjectives in a visual-linguistic context For example, one study proposed a cross-modal mapping between visual and word representations where the objective is to learn a function that assigns adjective labels to visual inputs [24], and more recently other researchers were able to learn a linear mapping that predicts an adjective descriptor given a visual input based on chaining [34]. Neural network models extend these works in image caption generation, which generally produces descriptive captions containing adjectives for a given query image [63]. Of course, one major limiting factor of these cross-modal approaches is that they may not be appropriate when dealing with relatively abstract, non-visual adjectives. All the work we are aware of either addresses change in language from a linguistic view, or presents a computational approach to assessing adjective-noun plausibility. To the best of our knowledge, there has been no prior work on formulating historical adjective extension as a computational

problem.

2.3 Changes in adjective use

Scholars have also investigated how the senses of adjectives and general set of nouns that they modify have changed over time. The most notable work is by linguist Joseph Williams, who proposed groups of adjectives transfer from one domain to another [62]. For instance, adjectives originally intended to describe touch perceptions have since extended to describe color (e.g., *warm cup* $\longrightarrow$ *warm color*). Similarly, adjectives originally intended to describe color have come to describe sound (e.g., *clear blue* $\longrightarrow$ *clear voice*), and vice-versa (e.g., *quiet room* $\longrightarrow$ *quiet blue*). Williams hypothesized that, in general, the domain transfer of adjectives follows general rules which he outlined in his paper. As this was a theoretical study, Williams only considered a select set of adjectives to which his principles apply, and so whether or not these rules generalize to all adjectives in English is still an open question.

It's unclear whether chaining plays a role in adjective domain transfer, but evidence from other areas of language use suggests chaining mechanisms play a big role in other types of transfer. More specifically, physical actions can turn to metaphors (e.g., to physically *grasp* an object $\longrightarrow$ to mentally *grasp* an idea) via chaining [65]. This recent work suggests adjective extension may indeed adhere to the same chaining principles, and makes our approach a promising research direction.

Chapter 3

Computational Models

In this chapter, we formalize hypotheses about semantic chaining presented over the past few decades into probabilistic models. These include the exemplar, prototype and k-nearest neighbors models and they all require some notion of semantics. We also describe a Bayesian framework under which our probabilistic models operate, including the use of a type-based prior. Lastly, we discuss how to learn parameters for our models and how we incorporate semantic information that is highly relevant to the predictive task.

3.1 The computational problem

We cast adjective extension as a temporal categorization problem. Given a noun n^*, our models seek to predict which adjectives $a \in \mathcal{A}$ are most appropriate for n^*, where a is drawn from a finite set of adjectives $\mathcal{A}$ that we consider. More formally, we are given two primary sources of information at time t and would like to predict the posterior probability that a will co-occur with n^* at time $t + \Delta$, denoted henceforth as $p(a|n^*)^{(t+\Delta)}$. The first source of information is a likelihood $p(n^*|a)^{(t)}$ that reflects (i) which other nouns paired up with a at time t, given as $\{n\}_a^{(t)}$, and (ii) n^*'s semantic relationship to every noun that paired with a at time t. The second is a prior probability $p(a)^{(t)}$ for adjective a and it tells us how likely a is to be re-used in a novel adjective-noun pairing at time $t + \Delta$. We combine the type-based prior with the likelihood to obtain the posterior probability of observing adjective a co-occur

with n^* at time $t + \Delta$:

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

In the above formulation, each adjective $a \in \mathcal{A}$ is treated as a category and thus described as the collection of nouns that co-occurred with a at time t. The prior and likelihood are described in more detail in chapters 3.2 and 3.3.

3.2 Prior distribution

We formulate a prior $p(a)^{(t)}$ that tells us how likely adjective a is to pair with any noun at time $t + \Delta$ in the absence of more specific semantic information. For a given adjective a, its prior probability is simply the normalized count of the number of unique nouns that it paired up with at time t:

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

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 [1, 28]. This rich-get-richer process is also supported by work on how semantic networks grow through preferential attachment [57]. In simpler terms, *wild* is likely to accumulate more nouns that *frenzied* as it is used more frequently by English speakers.

This type-based prior serves as our baseline model when making adjective predictions for n^* at time $t + \Delta$ and so the posterior probability assumes a uniform likelihood: $p\left(a|n^*\right)^{(t+\Delta)} = p\left(a\right)^{(t)}$. Using this prior as the baseline effectively assumes that adjective category size is the only factor governing how adjectives will extend, and ignores all hypotheses about semantic chaining. All chaining models (discussed in the next section) make use of the type-based prior when computing posterior probabilities.

3.3 Likelihood function

In this section, we describe how to compute the likelihood term $p(n^*|a)^{(t)}$ in our computational framework. Our likelihoods generally assume that classifier choices rely on *similarity* relationships between

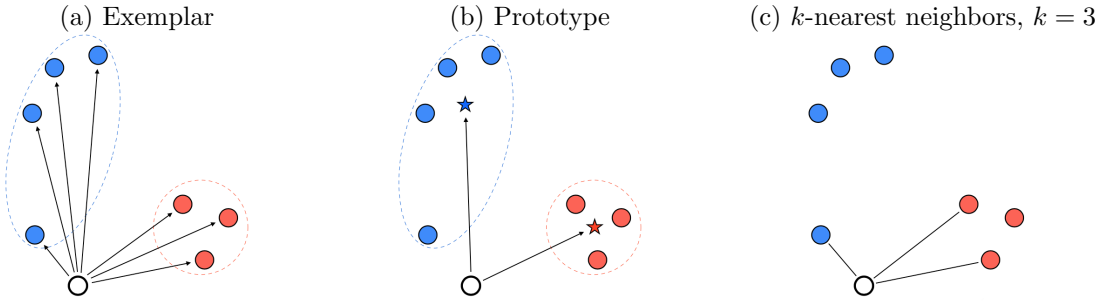


Figure 3.1: Illustration of the various chaining algorithms used to compute likelihood functions. The unshaded circle is the stimulus or the probe noun, red circles are nouns that have paired up with one particular adjective, and blue circles with another (although a noun may pair up with multiple adjectives).

noun n^* and other nouns that have paired with adjective a. We formally define the similarity between n^* and another noun n at time t as

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

where $d(\cdot,\cdot)$ is a distance metric [36, 52] and $\vec{\mathbf{v}}_{n^*}^{(t)}, \vec{\mathbf{v}}_{n}^{(t)}$ are semantic representations of nouns n^* and n respectively. All semantic representations of nouns are contained in a vector space over which the metric $d(\cdot,\cdot)$ is defined. Intuitively, similarity decreases exponentially as a function of distance. We use Euclidean distance as the metric in all our experiments. Figure 3.1 provides a visual illustration of our likelihood models.

3.3.1 Exemplar model

Exemplar theory [30] suggests that humans categorize a novel stimulus into one of finitely-many categories based on its degree of similarity with instances stored in memory. More specifically, for a given adjective a, the set of nouns that it co-occurred with at time t are its exemplars, and the stronger the relationship between n^* with a's exemplars, the more likely it is that n^* will pair with a at time $t + \Delta$. We consider a novel stimulus n^* and the goal is to obtain the appropriate likelihood $p(n^*|a)^{(t)}$. n^*'s degree of similarity to each of a's exemplars $n \in \{n\}_a^{(t)}$ determines the extent to which a is applicable to noun n^*:

$$\frac{1}{\left|\{n\}_a^{(t)}\right|} \sum_{n \in \{n\}_a^{(t)}} \text{sim}(n^*, n).$$

Note that we divide the exemplar sum by the the number of terms we sum over $\left|\{n\}_a^{(t)}\right|$ to control for category size. Exemplar theory takes semantic neighborhood density into account as nouns that are

closer in semantic space to n^* tend to dominate the likelihood.

In different categorization settings, we may want to control how sharply similarity declines as a function of distances in semantic space. In the psychological literature, the GCM is the most popular exemplar model as it introduces a kernel parameter $h^{(t)}$ (specific to time t) which controls the steepness of the similarity function [37]. This gives our exemplar likelihood

$$p\left(n^*|a\right)^{(t)} \propto \frac{1}{h^{(t)}\left|\{n\}_a^{(t)}\right|} \sum_{n \in \{n\}_a^{(t)}} \exp\left(-\frac{d\left(\vec{\mathbf{v}}_{n^*}^{(t)}, \vec{\mathbf{v}}_n^{(t)}\right)^2}{h^{(t)}}\right)$$

and we can recover the original version of the exemplar model by simply setting $h^{(t)} = 1$.

In this formulation, we also use $h^{(t)}$ as a normalizing term as this method of computing the likelihood $p\left(n^*|a\right)^{(t)}$ is effectively the same as performing kernel density estimation in semantic space [2, 41, 50]. More details about learning the kernel parameter are described in Chapter 3.4.

3.3.2 Prototype model

Prototype theory offers a different view on categorization than exemplar theory. Prototype theory, motivated mainly by Eleanor Rosch [48] with recent advancements in few-shot learning [55], suggests each category has a *prototype representation* which (i) humans associate with the category, and (ii) is representative of the category's exemplars. This implies that the strength of association between a stimulus noun n^* and adjective a's prototype representation determines the likelihood probability of n^* pairing up with a at time $t + \Delta$. Researchers have proposed multiple ways to compute the prototype $\vec{\mathbf{p}}_a^{(t)}$ for adjective a, but we follow a straightforward approach by simply computing the arithmetic mean of all exemplars $n \in \{n\}_a^{(t)}$ in semantic space [47]:

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

n^*'s degree of similarity to $\vec{\mathbf{p}}_a^{(t)}$ determines the likelihood probability:

$$p\left(n^*|a\right)^{(t)} \propto \exp\left(-\frac{d\left(\vec{\mathbf{v}}_{n^*}^{(t)}, \vec{\mathbf{p}}_a^{(t)}\right)^2}{h^{(t)}}\right)$$

where we once again fit a kernel parameter $h^{(t)}$ just as in our exemplar likelihood. This formulation makes use of the same similarity function (with the learned kernel parameter) as the exemplar model. However, unlike the exemplar model, here the likelihood is based on n^*'s proximity's to a's centroid and therefore doesn't take semantic neighborhood density into account.

3.3.3 Progenitor model

Theorists have argued as to whether category prototypes change once categories accumulate new exemplars. Some believe that the prototypical representation for a category remains fixed and hence the growth of the category radiates outwards from $\vec{\mathbf{p}}_a^{(t_0)}$ (where t_0 is the base time). The progenitor model is simply a variant of the prototype model where the prototype representations for adjective a remains "static", i.e., $\vec{\mathbf{p}}_a^{(t)} = \vec{\mathbf{p}}_a^{(t_0)}$ for all $t \geq t_0$.

3.3.4 k-nearest neighbors model

The basic idea behind distributional semantics is that examples within proximity of each other in semantic space exhibit similar properties [16]. The k-nearest neighbors (k-NN) model builds on this intuition along with recent work on distance-based representations in neural networks [19, 61]. The probability of n^* pairing up with a at time $t + \Delta$ is directly proportional to how many of n^*'s k-nearest nouns $n_1, n_2, \ldots, n_k$ paired with a at time t.

These nearest nouns are chosen based on semantics. Formally, let $\mathcal{N}_{\mathcal{A}}^{(t)}$ be the the finite set of nouns that we consider at time t (further described in Chapter 4.3) and let $\mathcal{S}_i \subseteq \mathcal{N}_{\mathcal{A}}^{(t)} \setminus \{n^*\}$ be any k-sized subset of $\mathcal{N}_{\mathcal{A}}^{(t)} \setminus \{n^*\}$ indexed by i. Given our choice of semantic space, define

$$\mathcal{S}_{k\text{NN}} = \arg\min_i \sum_{n \in S_i} d\left(\vec{\mathbf{v}}_{n^*}^{(t)}, \vec{\mathbf{v}}_n^{(t)}\right)$$

and enumerate the nouns in $\mathcal{S}_{k\text{NN}}$ as $n_1, n_2, \ldots, n_k$ (possibly randomly). Now that we've picked the k-nearest nouns to n^*, the k-NN likelihood in a Bayesian framework is

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

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-nearest nouns "voting" for which adjective they appeared with at time t. As it's highly likely that a

proximity noun n_j (where $j \leq k$) co-occurred with multiple adjectives, it votes multiple times—once for each of those adjectives. These votes ignore raw corpus co-occurrence counts, and instead are binary (i.e., type-based). This formulation can be interpreted as a somewhat "hard version" of the exemplar model as k is a discrete analog of the kernel parameter $h^{(t)}$. We report $k = 1$ and $k = 10$ in our experiments.

3.4 Kernel parameter estimation

We now describe our methodology for learning the kernel parameter $h^{(t)}$ used in both the exemplar and prototype models. The objective is to learn this kernel parameter using all the knowledge we have about historical and current adjective-noun pairings and the usage of nouns up to and including time t. By splitting our attested adjective-noun pairs into those that emerged at time $t - \Delta$ or before and those that emerged at exactly time t, we optimized precision (predictive accuracy) on the former set of pairings to predict the latter. This essentially allows us to treat novel pairs at time t as "validation data". More specifically, for any noun n, we define the function $f_n : \mathbb{R} \rightarrow \mathcal{A}^{m_n}$ which computes the the predicted posterior distribution and retrieves the adjective predictions for noun n given a kernel parameter $h^{(t)} \in \mathbb{R}$. Here, m_n is the number of adjectives $a \in \mathcal{A}$ that n first paired with at time t. We then perform the kernel parameter estimate as

$$\widehat{h}^{(t)} = \arg\max_{h} \frac{\sum_n \left| f_n(h) \cap \mathcal{J}_n^{(t)} \right|}{\sum_n m_n}$$

where $\mathcal{J}_n^{(t)}$ is the set of m_n adjectives that actually form novel adjective-noun pairs with n at time t. In the language of precision and recall, $f_n(h)$ gives the set of retrieved positives for a given kernel value h and $\mathcal{J}_n^{(t)}$ is the set of true positives. The estimated kernel parameter $\widehat{h}^{(t)}$ is then used to predict novel adjective-noun pairs at time $t + \Delta$ for all nouns. We learned a separate $h^{(t)}$ for each of the exemplar and prototype likelihoods at time t.

We chose to optimize the precision score as it reflects the extent to which our predictive models can accurately predict novel adjective-noun pairs. Of course, optimizing any model based on precision requires access to m_n (the number of new adjectives that pair with noun n at time $t+\Delta$) and it may seem odd for a model to know this quantity, but we treat this value as a cutoff. Our predictive models rank-order all adjectives that previously did not appear with n based on their predicted posterior probability and the top m_n such adjectives are compared to the emergent adjectives that pair with n at time $t + \Delta$. Since we are predicting a posterior categorical distribution over adjectives, it's also worth considering

metrics to evaluate the shape of the distribution itself, and we later go into more details about these approaches. These include using a maximum a posteriori estimate of the kernel parameter based on observed extensions (see Appendix D) as well as both optimizing and evaluating the Jensen-Shannon divergence between the predicted posterior and empirical distributions (see Appendix C).

3.5 Semantic space

The exemplar, prototype, and k-NN models all rely on some notion of semantics. More specifically, when estimating the likelihood term $p(n^*|a)^{(t)}$, we require meaningful representations of both n^* and each $n \in \{n\}_a^{(t)}$ to incorporate semantic similarity into these models. Word2Vec [31], and word embeddings based on distributional semantics in general, are a straightforward way to represent semantics and understand the uses of various nouns.

However, we need to account for the fact that our models are sequential and only observe co-occurrence statistics up to time t when making predictions for novel adjective-noun pairs at the next time step. Vector space models of distributional semantics are learned based on a word's co-occurrence distribution, thus word embeddings that have already "seen" adjective-noun combinations that emerged after time t are inappropriate for our predictive model as they peek ahead in time. Instead, we turn to diachronic Word2Vec embeddings [15]: these word embeddings are trained using a skip-gram model just as Word2Vec, however are time-sensitive. At each time t, the diachronic Word2Vec embedding for each noun is based solely on its co-occurrence statistics at time t, and all past and future co-occurrences are ignored. For instance, the uses of certain words such as *gay* and *awesome* over 100 years ago have almost no resemblance to their uses today as they have changed, and diachronic Word2Vec embeddings capture exactly that. Hence, the predictions made by our predictive models don't peak ahead at future adjective-noun pairings and are in a sense zero-shot. We used diachronic Word2Vec embeddings trained on a corpus of text written at time t when predicting the behavior of nouns with respect to their adjective pairings at time $t + \Delta$.

Chapter 4

Historical Data of Adjective Use

In this chapter, we describe a large corpus of historical adjective-noun pairings which we used as the basis for our study. This corpus comprises passages of written English over the past two centuries. We also describe three different sets of adjectives against which we test our computational models and how these sets were obtained. All data and code from our analyses are publicly available[1].

4.1 The Google books corpus

The Google books corpus [27] contains transcriptions of various pieces of text (mostly books) written over the past few centuries. Within the entire corpus, the English All (EngAll) corpus accounts for 8.5×10^{11} tokens and roughly 6% of all books ever published in English. The vastness of the EngAll corpus across various points in time is likely to reflect how written English has changed over the past few centuries, and since written and spoken language are generally tied together, we suspect this corpus as a whole is a good approximation of (English) language use over large periods of time. Words in this corpus are tagged with their part-of-speech (POS) and the year when the book in which they occur was published, and this is particularly useful as we are interested in adjectives, nouns, and when they co-occurred. As writers of English have changed their usage over the past few centuries (see Figure 1.1 for an example), we suspect that studying adjective-noun co-occurrences using the Google books corpus, and more specifically the EngAll corpus, is ideal for testing hypotheses about language change. We restrict our analysis to books published between 1800 and 2000.

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

4.2 Three adjective sets

In this section, we present three adjective sets $\mathcal{A}$ and their historical co-occurrences with nouns. This is to get a representative view of adjectives and to ensure our hypothesis is agnostic to choice of groups of adjectives. We evaluated our predictive models against each of these three adjective sets, and these sets are described in the following subsections.

4.2.1 Frequent adjectives

Our first set contains 200 frequently-used adjectives that cover a broad scope of descriptions. To construct this set, we first collected pre-trained Word2Vec embeddings[2] of all adjectives in WordNet [32]. Note that these are not diachronic Word2Vec embeddings, and instead reflect the modern use of English. Next, we performed 20-means clustering to group the adjectives based on semantics. Finally, we sampled 10 adjectives from each cluster based on frequency in the corpus across all times. That means that if adjective a was grouped into cluster $\mathcal{C}$, it was sampled with probability $f_a / \sum_{a' \in \mathcal{C}} f_{a'}$ where f_a is the raw occurrence frequency of a in the EngAll corpus between 1800 and 2000. For the remainder of this thesis, we refer to this set of adjectives as Frq-200.

4.2.2 Random adjectives

Frq-200 contains some of the most prominent adjectives in the English lexicon, but to test if our hypothesis about semantic chaining extends to a more general set of adjectives, we constructed $\mathcal{A}$ to contain 200 random adjectives that also cover a broad scope of descriptions. To construct this set, we followed the same protocol as with Frq-200 all while replacing frequency-based sampling with random sampling. That is, for a given cluster of adjectives $\mathcal{C}$, 10 adjectives were chosen via uniform sampling. We will refer to this set of adjectives as Rand-200. A brief comparison of adjectives in Frq-200 and Rand-200 is given in Table 4.3 and an exhaustive comparison is provided in Appendix A.

4.2.3 Synaesthetic adjectives

Both Frq-200 and Rand-200 are novel in the sense that we present them as new groupings of adjectives against which to evaluate our predictive models. However, linguists have previously studied specific adjectives and how they extend over time to describe different objects, such Joseph Williams' synaesthetic adjectives (discussed in Chapter 2.3). These are a pre-defined group of 65 adjectives[3] that have

[2] Available here: `https://code.google.com/archive/p/word2vec/`.

[3] Syn-65 actually contains 61 unique adjectives, and we go into detail about this adjective set in Appendix A.

historically extended between different sensory domains to describe various objects while exhibiting a pattern of regularity (e.g., touch to color: *warm cup* → *warm color*). We will refer to this set as Syn-65.

4.3 Nouns

To test our hypothesis about how adjectives behave and extend over time, we needed to pick a relevant set of nouns which capture almost all the uses of all adjectives in any of our adjective sets. For each adjective set $\mathcal{A}$ described in the last section, we followed a simple protocol to construct a set of relevant nouns that included (i) the most commonly paired nouns, and (ii) nouns that emerged after the base time t_0 (the base time is discussed in Chapter 5.1), both with respect to $\mathcal{A}$. As all adjective sets are almost mutually exclusive of each other, our noun sets differ for each $\mathcal{A}$.

Our protocol for constructing a set of nouns for adjective set $\mathcal{A}$ at time t, henceforth $\mathcal{N}_{\mathcal{A}}^{(t)}$, is as follows. First, we picked a base set of nouns by picking ones that co-occurred most frequently with adjectives in $\mathcal{A}$. That is, we scored each noun n based on its type-based frequency with all adjectives in $\mathcal{A}$ across all time periods:

$$\sum_{t} \sum_{a \in \mathcal{A}} \mathbb{1}\left[n \in \{n\}_{a}^{(t)}\right].$$

Based on this pseudo-metric, we selected the top 5,000 nouns as the base set for $\mathcal{A}$. Next, we deemed a noun to be *emerging* if and only if it first co-occurred with any $a \in \mathcal{A}$ at some time beyond the base time t_0. We picked the top 500 emerging nouns at each time t by ranking nouns that emerged at that time based on their type-based co-occurrence frequency with adjectives in $\mathcal{A}$. Consequently, $\mathcal{N}_{\mathcal{A}}^{(t)}$ grew as t increased. Letting $\mathcal{E}_{\mathcal{A}}^{(t+\Delta)}$ be the set of top 500 nouns that emerged with respect to $\mathcal{A}$ at time $t+\Delta$, the growth of our noun set is formally given by the recursive relation

$$\mathcal{N}_{\mathcal{A}}^{(t+\Delta)} = \mathcal{N}_{\mathcal{A}}^{(t)} \cup \mathcal{E}_{\mathcal{A}}^{(t+\Delta)}.$$

Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
Asian	*Hungarian*	*polite*	*chatty*	*warm*	*chilly*
Christian	*Thai*	*intelligent*	*unorthodox*	*dense*	*watery*
American	*Cornish*	*passionate*	*amiable*	*dry*	*fertile*
European	*Catalan*	*energetic*	*communicative*	*tropical*	*drizzling*

Table 4.1: A comparison of some adjectives in Frq-200 and Rand-200 grouped according to the cluster they were originally drawn from. Notice that the clusters align semantically, however the adjectives in Frq-200 are more prominent in the English lexicon than those in Rand-200.

Chapter 5

Experiments and Results

Having described our methodology in previous chapters, we now share results from testing our models against historical data of adjective use. In particular, in this chapter we discuss our experimental protocols and model performance. We also share results from additional studies that show adjectives don't extend randomly, but rather in accordance with the principle of cognitive economy. These results generally support our chaining hypothesis and suggest semantic chaining plays a significant role in adjective extension.

5.1 Experiment details

We tested our models and evaluated their predictive accuracy on each of the adjective sets detailed in Chapter 4.2 using the probabilistic formulation we described earlier. At each point in time t, each model was tasked with inferring which adjectives $a \in \mathcal{A}$ would pair with a given noun n^* at time $t + \Delta$. We divided our time bins into decades, i.e., $\Delta = 10$ years, and hence made predictions regarding how adjectives would extend in future decades. Although the EngAll corpus contains word counts between 1800 and 2000, we chose our base decade to be $t_0 = 1840$s. Note that despite this choice of t_0, we still count adjective-noun pairings as far back as 1800 and this gives all pairings between 1800 and the end of decade t, inclusive. Emergent adjective-noun pairs in the 1860s were the first against which we tested our predictive models (since we used historical data and embeddings from the 1840s to learn kernel parameters by treating the 1850s as a validation decade), and those in 1990s were the last.

Before we discuss more details about training our probabilistic models, we first turn to how we formally define the co-occurrence of adjective a with noun n^* as we have hand-waved this phrase in all previous chapters. Indeed, there are multiple ways to formally define how a and n^* co-occur, and one

straightforward account is their first raw co-occurrence (of the format "$a\_\text{ADJ}\ n^*\_\text{NOUN}$") in the EngAll corpus as marked by POS tags. A potential drawback of this definition is susceptibility to noise in the corpus, and to deal with this, we applied a threshold T to the number of co-occurrences between a and n^* in decade t for them to have formally co-occurred in decade t. In all our experiments, we used the threshold value $T = 2$ with the reason for this seemingly-arbitrary value being that if a and n^* co-occur at least twice, the chances of noise generating these observations is probably lower than just one co-occurrence, while we also do not want to ignore would-be novel pairs by applying a very high threshold. Given our threshold T, we say that adjective a and noun n^* co-occur for the first time (and are an emergent pair) in decade t if and only if the following two criteria are met:

1. a and n^* co-occur at least $T = 2$ times during decade t, and

2. during any decade $t' < t$, a and n^* never co-occurred at least $T = 2$ times.

It's worth noting that this definition of co-occurrence, a particular noun n^* may have "emerged" in different decades with respect to choice of adjective set. For instance, the first co-occurrence of n^* with any adjective $a \in \mathcal{A}$ may have been in decade t, in which case $n^* \in \mathcal{E}_{\mathcal{A}}^{(t)}$, yet the first co-occurrence of n^* with any adjective $a' \in \mathcal{A}'$ (i.e., a different adjective set) may have been in some later decade such as $t + \Delta$, in which case $n^* \notin \mathcal{E}_{\mathcal{A}'}^{(t)}$.

Now that we have described what it means for an adjective and noun to co-occur, we can return to discussing kernel parameter optimization. We learned individual kernel parameters for each of the exemplar and prototype models during each decade (while the progenitor model simply borrowed the first kernel parameter learned by our prototype likelihood) using the Nelder-Mead simplex method [35] and the exact training procedure is described in Chapter 3.4. When predicting novel adjective-noun pairs in decade $t + \Delta$, we first learned the kernel parameters through a training phase: we predicted the observed pairs that first emerged in decade t using the historical data up to decade $t - \Delta$ and the diachronic Word2Vec embeddings in decade t. These kernel parameters were then used to predict novel adjective-noun pairs that emerged in decade $t+\Delta$, and the models were augmented to now observe novel pairings in decade t.

Given a noun n^*, each model's output was a categorical distribution $p\left(a|n^*\right)^{(t+\Delta)}$ over adjectives in $\mathcal{A}$. The adjectives were rank-ordered based on posterior probability and we report aggregate precision scores for all predictions that a model made in decade $t + \Delta$. This means that if n^* formed m novel adjective-noun pairs in decade $t + \Delta$, we considered the top m adjectives from our rank-ordering. However, since we're only concerned with predicting novel adjective-noun pairs that first emerge in decade $t + \Delta$, we ignored adjectives $a \in \mathcal{A}$ which have paired with n^* at any time up to and including t. Given this

criterion, we rewrite the posterior probability $p(a|n^*)^{(t+\Delta)}$ that we formulated in Chapter 3.1 such that it ignores previously-attested adjective-noun pairs:

$$p\left(a|n^*\right)^{(t+\Delta)} \propto p(n^*|a)^{(t)} p\left(a\right)^{(t)} \times \mathbb{1}\left[n \notin \{n\}_a^{(\leq t)}\right]$$

where $\{n\}_a^{(\leq t)}$ uses similar notation that we introduced in Chapter 3 and gives all nouns that co-occurred with a in at least one decade up to t.

Furthermore, we evaluated our predictive models in two ways. First, for a given noun n^*, we considered the set of true positives to be adjectives in $\mathcal{A}$ that first co-occurred with n^* specifically in decade $t + \Delta$. This means if some adjective a first appeared with n^* in some future decade $t' > t + \Delta$ and the model predicted a, it was counted as an incorrect prediction. As we learned kernel parameters for the exemplar, prototype, and progenitor models by maximizing the predictive accuracy on attested pairings specifically in the following decade, this is an obvious choice for set of true positives. Second, it doesn't seem entirely correct to penalize the model's predictive performance because it predicted some adjective that co-occurred with n^* eventually, even though not exactly in decade $t + \Delta$. For this reason, we relax the set of true positives to include any adjective that first co-occurred with n^* in any future decade $t' > t$. We report model predictive accuracy using both sets of true positives, however we didn't learn a separate set of kernel parameters for the latter set. In the rest of this chapter, we use these experimental protocols described in this section and report our results.

5.2 Predictive model performance

We tested all models and evaluated their predictive accuracy using the methodology described in the previous section. Aggregate precision accuracy (when considering emergent adjective-noun pairs in decade $t + \Delta$ only) is reported in Figure 5.1. The rank-ordering of the models remains consistent across the Frq-200, Rand-200, and Syn-65 adjective sets. The exemplar model obtained the highest precision accuracy and was closely followed by the 10-NN model. The difference between the exemplar and 10-NN model is almost negligible. This makes sense as the exemplar likelihood is essentially a Gaussian mixture model in semantic space and the kernel parameter controls how fast the similarity function decays, and so the exemplar model can pick how many neighbors to pay attention to, whereas any k-NN model is more stringent since k is fixed. Figure 5.3 provides a concrete example of chaining and the exemplar model. The 1-NN model was significantly worse than all other models, including the type-based prior. The prototype model followed closely behind the exemplar and k-NN models. The progenitor model,

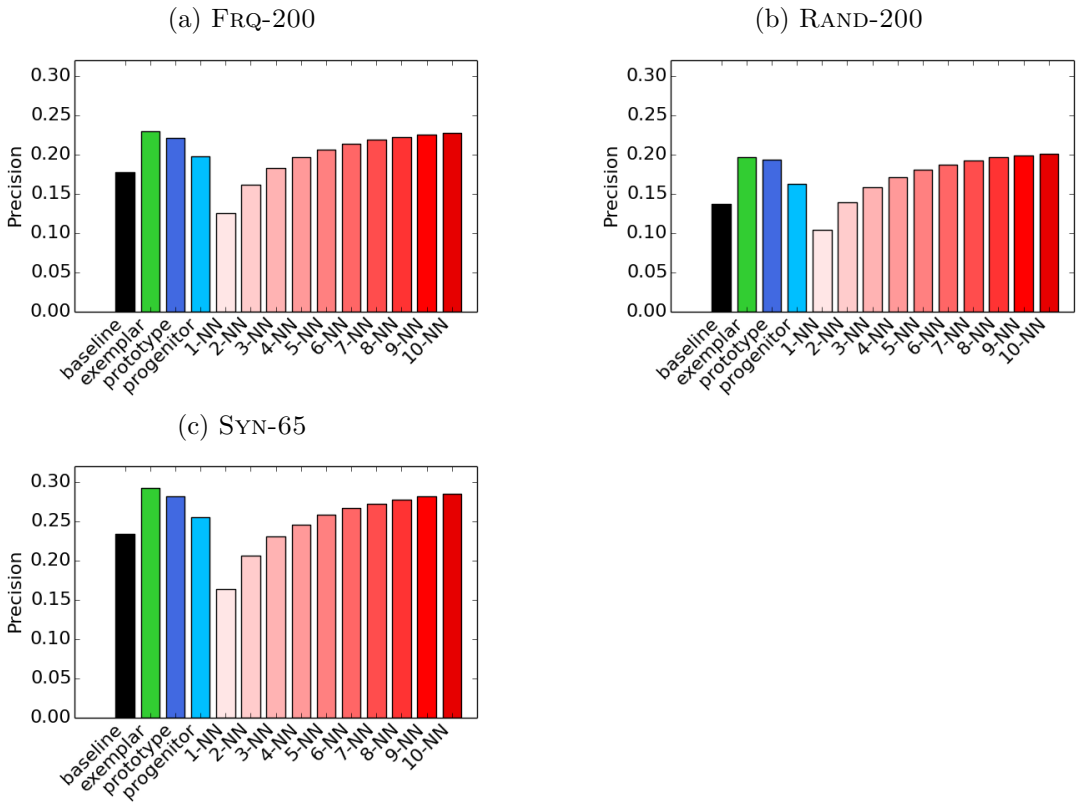


Figure 5.1: Aggregate precision accuracy for all models (including k-NN from $k = 1$ to $k = 10$) across all decades on each of the three adjectives sets.

a variant of the prototype model with "static" prototypes determined in the base decade t_0, became considerably worse than the prototype model with time. During later decades, the progenitor model performed even worse than the baseline. This relationship between the prototype and progenitor models that we observe here indicates that the context distribution of a affects which nouns n will pair with a in decade $t + \Delta$ to a great extent. This evidence suggests that adjective prototype representations are constantly changing and need to adapt to changes in semantics, and 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. Model predictive accuracy is also reported on a per-decade basis in Figure 5.2 and reflects the same aggregate trend from Figure 5.1. These results are consistent across all three adjective sets We also report some examples of model predictions in Table 5.1.

We also report accuracy where we relax the set of true positives for a given noun n^* to be all adjectives in $\mathcal{A}$ that formed novel adjective-noun pairs with n^* in any future decade $t' > t$ (see plots on the right-hand side of Figure 5.2). The rank-ordering for the top three performing models (exemplar, prototype, and 10-NN) wasn't always consistent between adjective sets, however the gap is quite small, suggesting all three are equally powerful for this method of evaluation. The same trend of the progenitor,

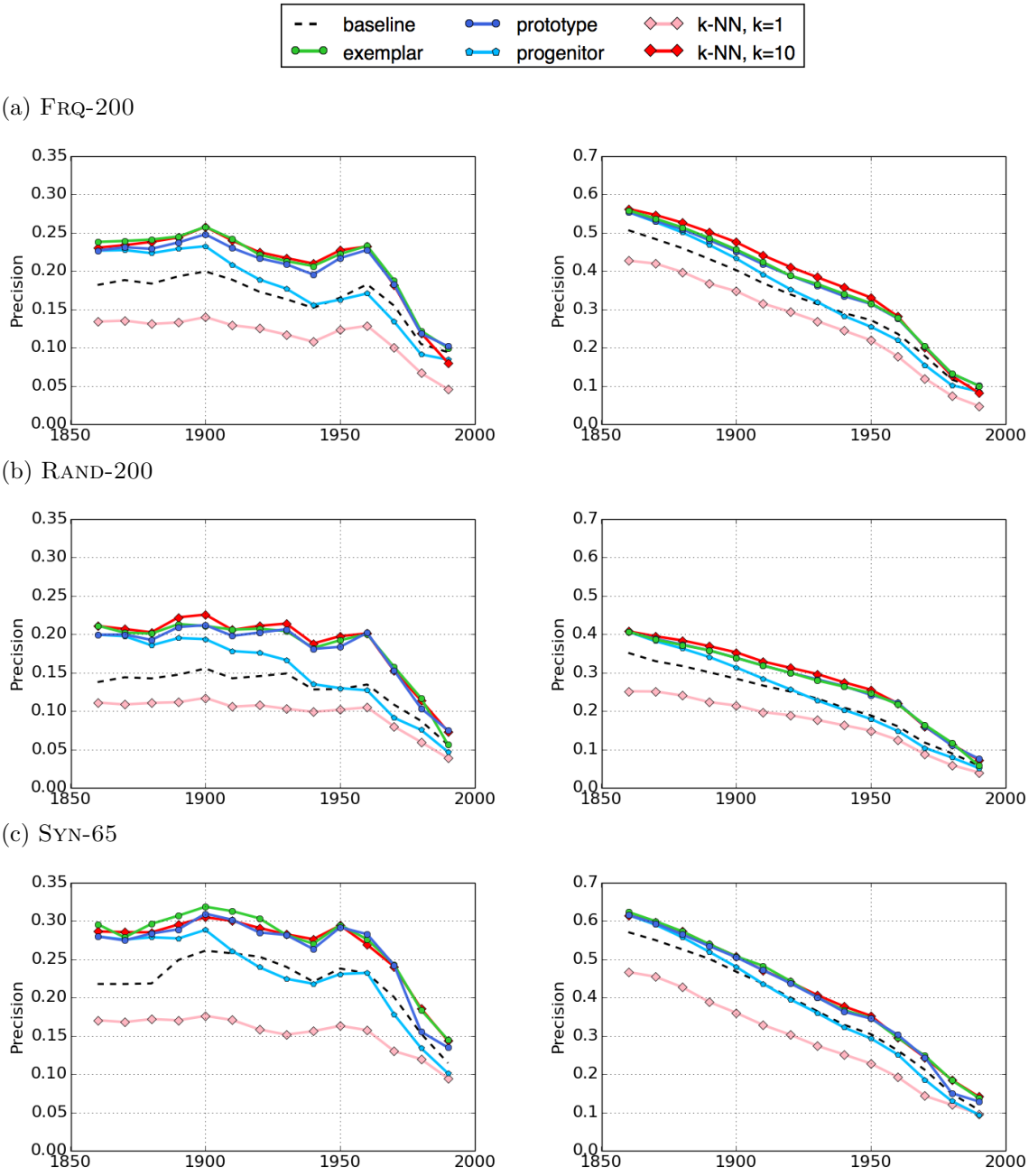


Figure 5.2: Model predictive accuracy on the Frq-200, Rand-200, and Syn-65 adjective sets. **Left:** Predictive accuracy when only novel adjective-noun pairs in the following decade are considered. **Right:** Predictive accuracy when all future adjective extension are considered. See Appendix B for an argument as to why the curves generally decrease with time.

baseline, and 1-NN model being the worst three performers followed. When evaluating our models this way, we borrowed the same kernel parameters as when only considering novel adjective-noun pairs in decade $t + \Delta$. This is possibly why the exemplar model doesn't always outperform the 10-NN model in this case, and we suspect a new set of learned kernel parameters would only increase precision accuracy

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*, *pool*, *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	*cigarette*, 1920s
new adjectives	*different*, *odd*, *worn*, *scattered*, *illegal*, *wrong*
baseline prediction	*natural*, ***different***, *sufficient*, *extraordinary*, *moral*, *mental* (1/6)
exemplar prediction	***different***, *natural*, *warm*, *sufficient*, *solid*, *inner* (1/6)
prototype prediction	*warm*, ***different***, *top*, *natural*, *solid*, *circular* (1/6)
10-NN prediction	*natural*, *top*, *warm*, ***different***, *sufficient*, *conventional* (1/6)

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	*discrimination*, 1940s
new adjectives	*female*, *south*, *predictive*, *silent*, *dietary*, *deplorable*
baseline prediction	*roman*, *solid*, *brilliant*, *unknown*, ***silent***, ***female*** (2/6)
exemplar prediction	***female***, *roman*, *male*, ***silent***, *solid*, *passionate* (2/6)
prototype prediction	*roman*, *male*, ***female***, *exaggerated*, *bourgeois*, *everyday* (1/6)
10-NN prediction	***female***, *male*, *exaggerated*, *energetic*, *isolated*, *roman* (1/6)

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)

Table 5.1: Examples of model predictions on the Frq-200 adjective set. Adjectives in bold font indicate true positives retrieved by models. For each of the model predictions, the adjectives are ordered in decreasing order of predicted posterior probability. We specifically include predictions for nouns *cigarette*, *alcohol*, and *Vietnam* as the adjectives they first pair in various decades reflect sentiment (e.g., *social cigarette* in the 1880s versus *wrong cigarette* nearly half a century later) or historic events (e.g., *illegal alcohol* due to prohibition, *American Vietnam* due to the Vietnam war).

by a small amount.

Predictive accuracy generally decreased with time both when we considered novel adjectives that emerged with n^* exactly in decade $t + \Delta$ and any later date. We go into detail about this phenomenon in Appendix B and show how the number of novel adjective-noun left for each model to predict decreased

across all three adjective sets as time increases, and this made the predictive task harder.

Another interesting finding is falsely-reported adjective-noun pairings as a result of misidentified extractions. For instance, Table 5.1 shows how *more* co-occurred with *Vietnam* in a novel adjective-noun pairing, yet *more* should not be treated as an adjective. In fact, *more* accounts for roughly 3.5% of all adjective-noun co-occurrences between adjectives in Frq-200 and all nouns in WordNet—a significant amount since a uniform distribution of co-occurrences over Frq-200 would include *more* just 0.5% of all co-occurrences. This reveals how POS tags can incorrectly factor into our data since *more* is treated as an adjective both by WordNet and the Google Books corpus.

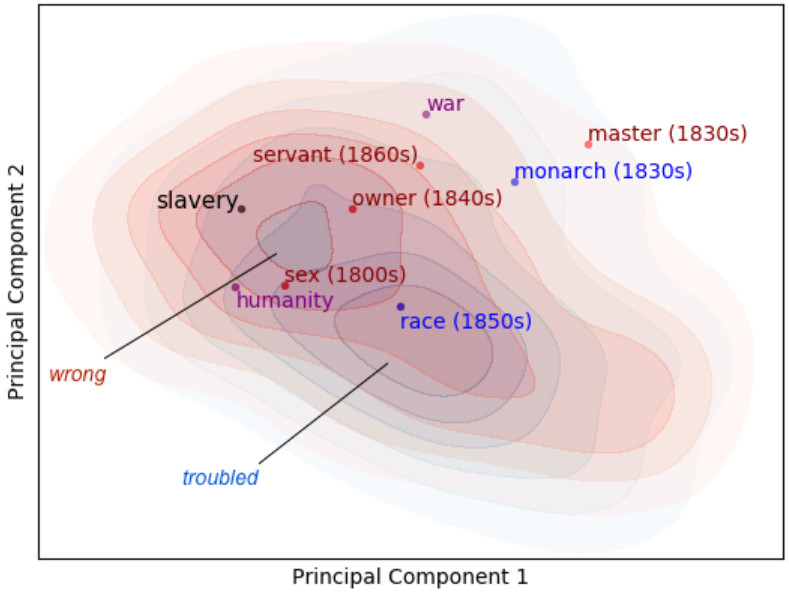


Figure 5.3: An illustration of chaining and more specifically the exemplar model. Both *wrong* and *troubled* compete to attract the noun *slavery* in the 1880s, prior to which *slavery* did not pair with either adjective. The exemplar model essentially forms a kernel density estimate for each adjective, shown here by the red contours for *wrong* and blue for *troubled*. Several nouns that helped to form each likelihood are also illustrated along with the decade in which they first paired with their respective adjectives (nouns labeled in purple previously paired with both adjectives). It's more likely that *wrong* will attract *slavery* than *troubled* will as the density function of the former assigns a higher probability to *slavery*, and indeed this gave rise to a novel adjective-noun pair in the 1880s: *wrong slavery*. We applied Principal Components Analysis to diachronic Word2Vec embeddings from the 1870s to obtain this figure.

5.3 Optimal k-nearest neighbors model

Our results from Chapter 5.2 led us to believe that chaining is a step in the right direction towards capturing the generative cognitive mechanisms that underlie adjective extension, and also that semantic neighborhood is an important contributing factor since the exemplar and 10-NN models achieved the highest predictive accuracy. We now explore roughly how large this semantic neighborhood is in terms of neighboring nouns. Despite that the 10-NN model largely outperformed the 1-NN model, it's clear that model performance will start to decrease after some value of k. This is because in a type-based k-NN setting (such as ours), model performance converges to the baseline as $k \to \left|\mathcal{N}_{\mathcal{A}}^{(t)}\right|$ (i.e., a type-based k-NN model with $k = \left|\mathcal{N}_{\mathcal{A}}^{(t)}\right|$ is the baseline by definition.

Furthermore, one intuition of the kernel parameter in the exemplar model (which we deem a soft-version of k-NN) is how many neighbors are "paid attention to" as the kernel parameter effectively controls the steepness of the similarity function. To investigate how many neighbors are relevant to predicting the adjectives that a noun n^* will later pair with, we performed a grid search to find the optimal k that yielded the greatest predictive accuracy. We found that $k \in [30, 60]$ generally yielded the best performance. Figure 5.4 illustrates the distribution of top-performing k values under both evaluation modes. Across all decades, the best performing k-NN model never exceeded 10-NN by more than 0.03 precision accuracy.

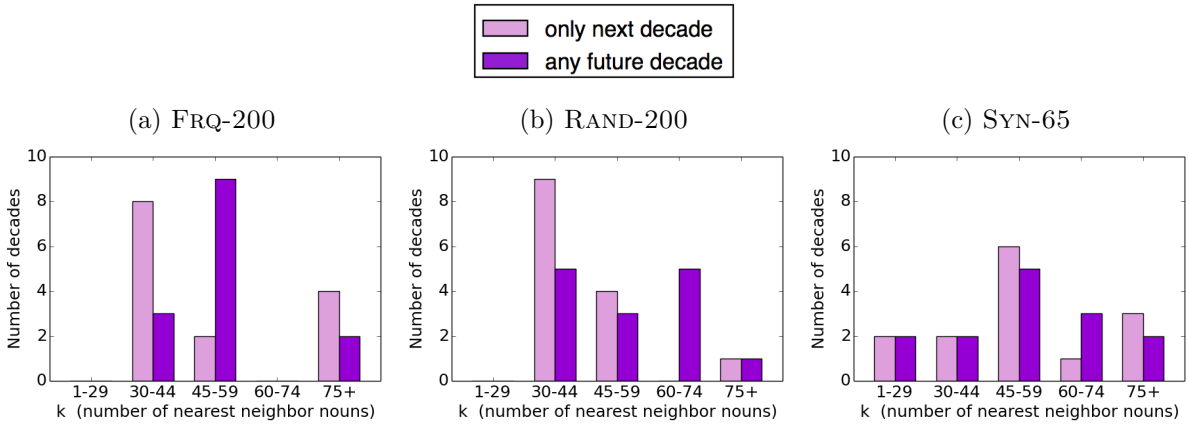


Figure 5.4: The distribution of optimal k values in a k-NN predictive model (based on precision) over all decades and across the three adjective sets when searching with step size 5.

5.4 The cognitive cost of adjective extension

Linguists have claimed that semantic chaining largely explains language change. We make the same hypothesis, as chaining facilitates extensions that are the cognitively "cheap" and this conforms to

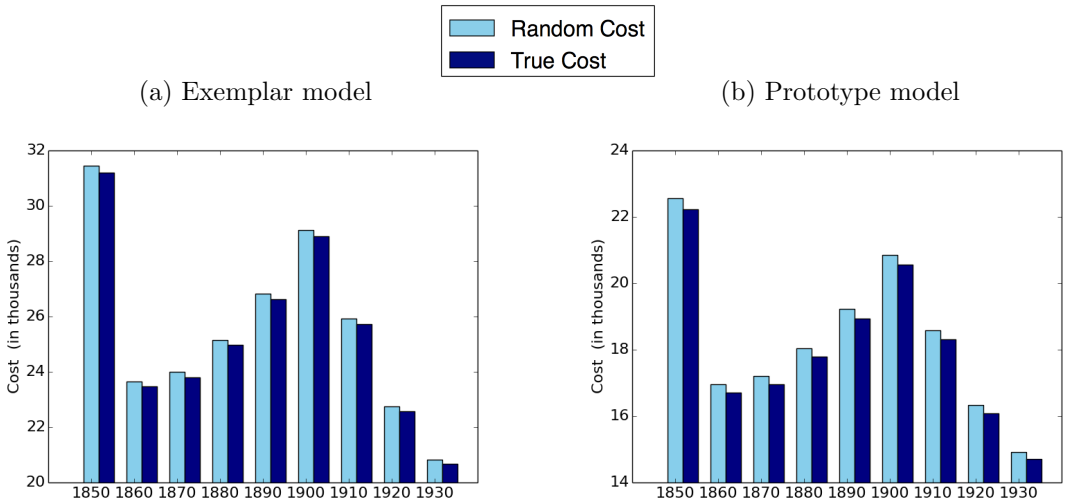


Figure 5.5: The "cognitive cost" of adjective extension for both the exemplar and prototype models on the Frq-200 adjective set. The true cost (light blue) is an empirical measurement based on each adjective's novel noun pairings in decade t over all $a \in \mathcal{A}$. The randomized cost (dark blue) is an average of 25 simulations where, if adjective a had actually formed m novel adjective-noun pairs in decade $t+\Delta$, then those m nouns were chosen randomly out of the set of nouns a hadn't previously paired with. Error bars are omitted here as any confidence interval around the mean random cost is too small to see.

the general principle of cognitive economy, which is the tendency for cognitive processes to minimize processing effort and resources. However, it's not clear how we can find empirical evidence that supports or rejects our hypothesis, especially since the underlying mechanism which explains adjective extension is not well-understood at neither the cognitive nor neural levels.

As our predictive models shed some light on the generative processes at play, we tried to measure the "cognitive cost" of adjective extension through our exemplar and prototype models versus a random process of extension, and this showed the extent to which our claim about adjective extension acting in accordance with the principle of cognitive economy is true. Note that we use the term "cognitive cost" loosely as it's not well-defined. As chaining primarily operates in a semantic space, we computed costs based on how far in that semantic space an adjective needed to extend to include a new noun. With respect to the exemplar model, we defined the cost of extending adjective a to noun n^* (which a previously never paired with) in decade $t+\Delta$ as

$$\text{cost}(a, n^*) = \frac{1}{\left|\{n\}_a^{(t)}\right|} \sum_{n \in \{n\}_a^{(t)}} d\left(\vec{\mathbf{v}}_{n^*}^{(t)}, \vec{\mathbf{v}}_n^{(t)}\right).$$

Similarly, we defined the same cost with respect to the prototype model as

$$\text{cost}(a, n^*) = d\left(\vec{\mathbf{v}}_{n^*}^{(t)}, \vec{\mathbf{p}}_a^{(t)}\right).$$

This way of measuring cost reflects how distances in semantic space grow with concepts that are more and more unrelated (in the distributional sense), a direct consequence of the learning procedure in Word2Vec. Results from our experiment are reported in Figure 5.5 and they suggest that adjective extension is by no means random, i.e., when an adjective extended to a noun, the distribution over nouns is far from uniform. While we acknowledge that our notion of "cheapness" is biased since its based solely on Word2Vec embeddings, we report our results with $p < 0.01$.

Chapter 6

Discussion

This thesis has provided a computational formulation of adjective extension, a large dataset of historical adjective-noun pairings along with their usages throughout the last 150 years, and an empirical evaluation of probabilistic models that recapitulate the extension of adjectives based on the idea of semantic chaining. When evaluating the predictive accuracy of our models, the exemplar model tends to outperform all others, followed closely by the 10-NN and prototype models. These models performed considerably better than the type-based prior that extends adjectives based solely on category size. This result was consistent throughout almost all decades into which we attempted to predict novel adjective-noun pairs, and across our three adjective sets.

Our evidence supports our hypothesis that semantic neighborhood density influences which novel adjective-noun pairs will emerge albeit not too strongly as our prototype model is able to perform nearly as well as our exemplar and nearest neighbors models in terms of predictive accuracy. The prototype model, unlike the exemplar and nearest neighbors models, does not account for semantic neighborhood density. Therefore, as chaining can take various forms, our results do not illustrate which mechanism of adjective extension is necessarily preferred, if any. However, these results do suggest that if a prototype model is central to adjective extension, then the prototype representations are constantly adapting to novel contexts since our progenitor models worsened with time as compared with our prototype model. Our work builds on previous efforts that suggests chaining algorithms are the computational underpinnings for language change in domains such as the extension of word senses, container names, and numeral classifiers. This thesis investigated whether those same chaining mechanisms apply to adjective extension, and we found that the way in which language changes with respect to adjective-noun composition generally follows the same principles uncovered in those other domains. Together with

previous work, our results imply that chaining plays a crucial role in explaining how language changes over time.

Our conclusion matches our hypothesis for the most part, yet our approach still has some limitations that are worth discussing and potentially addressing in future work. First, we specifically used a type-based approach when computing both the prior distribution and likelihood in our computational framework, but it's worth noting that both type- and token-based approaches can be employed towards categorization tasks such as ours. The latter counts token frequency rather than binary co-occurrence counts. In fact, token-based counts are generally more common with regards to cognitive modelling, but it's not clear which approach is better, and some work in the psychology literature has pointed towards type-based representations as being superior [43] while others have hinted token-based representations [4, 38] are better. Indeed, if token-based counts are more informative of adjective categories, then this would largely affect our predictive models. Second, chaining is primarily based on the notion of semantic similarity. One drawback of this general setup is that although chaining mechanisms may retrieve other stimuli that are similar to our query, plausibility is still ignored. That is, our implementation of chaining does not explicitly "perform a check" as to whether the adjective predictions for a given query noun give sensible pairings. This is perhaps why our models still make predictions such as *solid discrimination* (see Table 5.1) which is nonsensical with respect to any known sense of the adjective *solid*, and in fact never became attested. As adjectives are able to accumulate novel senses and uses, the set of feasible nouns they can pair with will vary, thus there is no clear solution to this problem. At present, we acknowledge this limitation but also suspect any potential solution that can disregard nonsensical composition will yield a significant increase in predictive accuracy.

As our work builds certain inductive biases into our predictive models based on what we know about chaining and category growth, future avenues should explore richer models for temporal prediction with a stronger and more informative set of biases. Here, we a discuss a few of these directions. First, the semantic representations that our models leverage largely impact their performance and predictions. In this thesis, we used Word2Vec, which captures the distributional semantics of words. However, the way we choose to represent semantics need not rely on the distributional hypothesis, and the optimal choice of semantic representation may be task-specific. A variant of Word2Vec which has been lauded for better overall representations is GloVe [42], based on global co-occurrence statistics, and non-distributional representations which may further contribute to a more informative likelihood function include, and are not limited to, binary sparse representations of words [11]. Next, just as we formulated semantic chaining in computational terms, we can also leverage other hypotheses about language change and incorporate them into our predictive models as inductive biases. For instance, the law of parallel change

states that related words tends to change in similar ways [25], and has already been investigated in other domains of language change [64]. As nouns can often be described by multiple adjectives that ultimately communicate the same meaning, the law of parallel change may also help guide our predictive models in adjective extension. Finally, as studies have shown that the visual world captures useful information about sensible adjective-noun pairings, can we incorporate a visual component into our predictive models? In the past, studies that have taken a cross-modal approach to adjective-noun composition have targeted adjectives that are relatively concrete, such as *blue* and *wooden*, and thus we face the greater challenge of making cross-modal models deal with abstract adjectives in the lexicon—if a cross-modal approach can offer any benefits at all.

To conclude, this thesis 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 how human cognition, coupled with external factors such as changes in culture and technological advances, shape innovate language use.

Bibliography

[1] John R. Anderson. The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429, 1991.

[2] F. Gregory Ashby and Leola A. Alfonso-Reese. Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39(2):216–233, 1995.

[3] Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, 2010.

[4] Lawrence W. Barsalou, Janellen Huttenlocher, and Koen Lamberts. Basing categorization on individuals and events. *Cognitive Psychology*, 36:203–272, 1998.

[5] Gemma Boleda, Marco Baroni, Louise McNally, and Nghia Pham. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 7th International Conference on Computational Semantics*, pages 35–46, 2013.

[6] Joan L. Bybee. Usage-based theory and exemplar representations of constructions. *The Oxford handbook of construction grammar*, 2013.

[7] Joan L. Bybee, Revere D. Perkins, and William Pagliuca. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press, 1994.

[8] Noam Chomsky. *Syntactic Structures*. Mouton, 1957.

[9] Katrin Erk. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 216–223, 2007.

[10] Katrin Erk, Sebastian Padó, and Ulrike Padó. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763, 2010.

[11] Manaal Faruqui and Chris Dyer. Non-distributional word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 464–469, 2015.

[12] Dirk Geeraerts. *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Oxford University Press, 1997.

[13] Emiliano Guevara. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, GEMS Workshop*, 2010.

[14] Amir Ahmad Habibi, Charles Kemp, and Yang Xu. Chaining and the growth of linguistic categories. *Cognition*, to appear.

[15] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, 2016.

[16] Zellig S. Harris. *Mathematical structures of language*. Wiley, 1968.

[17] Emmanuel Keuleers. *Memory-based learning of inflectional morphology*. PhD thesis, Universiteit Antwerpen, 2008.

[18] Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015.

[19] Gregory Koch. Siamese neural networks for one-shot image recognition. Master's thesis, University of Toronto, 2015.

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

[21] George Kingsley Lakoff. *Human behavior and the principle of least effort*. Addison-Wesley Press, 1949.

[22] Maria Lapata. 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*, pages 63–70, 2001.

[23] Maria Lapata, Scott McDonald, and Frank Keller. Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 30–36, 1999.

[24] Angeliki Lazaridou, Georgiana Dinu, Adam Liska, and Marco Baroni. From visual attributes to adjectives through decompositional distributional semantics. *Transactions of the Association for Computational Linguistics*, 3:183–196, 2015.

[25] Adrienne Lehrer. The influence of semantic fields on semantic change. *Historical Semantics: Historical Word Formation*, 29:283–296, 1985.

[26] Barbara Lewandowska-Tomaszczyk. Polysemy, prototypes, and radial categories. *The Oxford handbook of cognitive linguistics*, 2007.

[27] Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwantand Will Brockma, and Slav Petrov. Syntactic annotations for the google books ngram corpus. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 169–174, 2012.

[28] Yiwei Luo and Yang Xu. Stability in the temporal dynamics of word meanings. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 2018.

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

[30] Douglas L. Medin and Marguerite M. Schaffer. Context theory of classification learning. *Psychological Review*, 85(3):207–238, 1978.

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

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

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

[34] Tushar Nagarajan and Kristen Grauman. Attributes as operators: Factorizing unseen attribute-object compositions. In *Proceedings of the 15th European Conference on Computer Vision*, 2018.

[35] John A. Nelder and Roger Mead. A simplex method for function minimization. *Computer Journal*, 7(4):308–313, 1965.

[36] Robert M. Nosofsky. Luce's choice model and Thurstone's categorical judgment model compared: Kornbrots data revisited. *Perception & Psychophysics*, 37:89–91, 1985.

[37] Robert M. Nosofsky. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57, 1986.

[38] Robert M. Nosofsky. Similarity, frequency, and category representations. *Experimental Psychology: Learning, Memory, and Cognition*, 14:54–65, 1988.

[39] Sebastian Padó, Ulrike Padó, and Katrin Erk. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 400–4009, 2007.

[40] Gary B. Palmer and Claudia Woodman. Ontological classifiers as polycentric categories, as seen in shona class 3 nouns. In Martin Pütz and Marjolijn Verspoor, editors, *Explorations in Linguistic Relativity*, pages 225–149. John Benjamins Publishing Company, Amsterdam; Philadelphia, 2000.

[41] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 3(3):1065–1076, 1962.

[42] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[43] Amy Perfors, Keith Ransom, and Daniel J. Navarro. People ignore token frequency when deciding how widely to generalize. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2014.

[44] Janet B. Pierrehumbert. Exemplar dynamics: Word frequency, lenition and contrast. In Joan L. Bybee and Paul J. Hopper, editors, *Frequency and the emergence of linguistic structure*, volume 45, pages 137–157. John Benjamins Publishing Company, Amsterdam, 2001.

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

[46] Rachel Ramsey. *An exemplar-theoretic account of word senses.* PhD thesis, Northumbria University, 2017.

[47] Stephen K. Reed. Pattern recognition and categorization. *Cognitive Psychology*, 3:382–407, 1972.

[48] Eleanor Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233, 1975.

[49] Eleanor Rosch. Principles of categorization. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1978.

[50] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[51] Lauren A. Schmidt, Charles Kemp, and Joshua B. Tenenbaum. Nonsense and sensibility: Inferring unseen possibilities. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 2006.

[52] Roger N. Shepard. Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, 55(6):509–523, 1958.

[53] Royal Skousen. *Analogical modelling of language*. Kluwer Academic Publishers, 1989.

[54] Steven A. Slomand, Barbara C. Malt, and Arthur Fridman. Categorization versus similarity: The case of container names. *Similarity and Categorization*, pages 73–86, 2001.

[55] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017.

[56] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211, 2012.

[57] Mark Steyvers and Joshua B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29:41–78, 2005.

[58] Zhewei Sun, Richard Zemel, and Yang Xu. Slang generation as categorization. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, 2019.

[59] Eva M. Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science*, 41(2):102–136, 2017.

[60] Eva Maria Vecchi, Roberto Zamparelli, and Marco Baroni. Studying the recursive behaviour of adjectival modification with compositional distributional semantics. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 141–151, 2013.

[61] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016.

[62] Joseph M. Williams. Synaesthetic adjectives: A possible law of aemantic change. *Language*, 32(2):461–78, 1976.

[63] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Richard S. Zemel Ruslan Salakhutdinov, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[64] Yang Xu and Charles Kemp. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 2015.

[65] Yang Xu, Barbara C. Malt, and Mahesh Srinivasan. Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive Psychology*, 96:41–53, 2017.

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

[67] Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1263–1271, 2010.

Appendices

Appendix A

Adjective Sets

Here we present all adjectives used in our analysis, namely from the FRQ-200, RAND-200, and SYN-65 adjective sets. Adjectives in bold font are included in at least two of the three adjective sets. The first table gives the adjectives that constitute SYN-65, and we note two important details about this set. First, the set of synaesthetic adjectives proposed by Joseph Williams [62] actually contains 64 unique adjectives as *light* is repeated. Second, the Google Books corpus ties all tokens to words in WordNet, and since *acrid*, *aspre*, and *tart* (all synaesthetic adjectives) are not WordNet adjectives, we could not reliably measure their uses through time. For this reason, we excluded these from SYN-65 and have 61 adjectives in total, given here.

SYN-65						
acute	*cloying*	***dulcet***	*grave*	*light*	*quiet*	*sour*
austere	*coarse*	*dull*	*hard*	*little*	*rough*	*strident*
big	***cold***	*eager*	*harsh*	*loud*	*shallow*	*sweet*
bitter	*cool*	*empty*	*heavy*	*low*	*sharp*	*thick*
bland	*crisp*	*even*	*high*	*mellow*	*shrill*	*thin*
bright	*dark*	*faint*	*hollow*	*mild*	*small*	*vivid*
brilliant	*deep*	*fat*	*hot*	*piquant*	*smart*	***warm***
brisk	*dim*	*flat*	*keen*	*poignant*	*smooth*	
clear	***dry***	*full*	*level*	*pungent*	*soft*	

Next, we present the FRQ-200 and RAND-200 adjective sets. Since these two sets draw adjectives from identical clusters, we present the two adjective sets so we can easily compare adjectives drawn from same cluster between the two sets. See Chapter 4.2 for more details on how these sets were generated.

cluster 1 of 20		cluster 2 of 20		cluster 3 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
casual	*amiable*	*bare*	*contorted*	*sufficient*	*alterable*
eccentric	*chatty*	*curly*	*dainty*	*analogous*	*contemporaneous*
energetic	*communicative*	*eyed*	*furrowed*	*equal*	*reconcilable*
entertaining	*fiery*	*female*	*hale*	*calculable*	*chargeable*
enthusiastic	*fluent*	*feminine*	***horny***	*receivable*	*distributive*
forgiving	*guileless*	***horny***	*limber*	*derived*	*accessary*
glib	*lovable*	*male*	*sage*	*binding*	*lineal*
intelligent	*loyal*	*naked*	***skeletal***	*indirect*	*allotted*
passionate	*patriotic*	*pale*	*smoky*	*undivided*	*noncommercial*
polite	*unorthodox*	***skeletal***	*swaggering*	*eligible*	*classifiable*

cluster 4 of 20		cluster 5 of 20		cluster 6 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
cold	*chilly*	*algebraic*	*binary*	*blind*	*intact*
dense	***cold***	*conventional*	*biotic*	*impossible*	*irretrievable*
dry	*drizzling*	*discrete*	*crystalline*	*incomplete*	*malfunctioning*
eastern	*encroaching*	*electrical*	*fusible*	*isolated*	*obscure*
hardy	*fertile*	*microscopic*	*geometric*	*pregnant*	*overlooked*
northern	*funicular*	*multicellular*	*interfacial*	*scarce*	*powerless*
south	*homeward*	*predictive*	*modular*	*silent*	*unmarked*
tropical	*littoral*	*rotational*	*perceptual*	*submerged*	*unstable*
warm	*unincorporated*	*thermal*	*refrigerant*	*unknown*	*unstudied*
western	*watery*	*volcanic*	*stratified*	*unrelated*	*valueless*

cluster 7 of 20		cluster 8 of 20		cluster 9 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
appropriate	*complex*	*alien*	*antipodal*	*everyday*	*approaching*
balanced	*delighted*	*colonial*	*congruous*	*firm*	*descending*
basic	*foolproof*	*divine*	*dynastic*	*more*	*fiddling*
better	*grateful*	*heavenly*	*hierarchical*	*original*	*former*
different	*intensive*	*human*	*invariable*	*particular*	*intensifying*
natural	*knowledgeable*	*inner*	*overt*	*physical*	*probable*
positive	*livable*	*medieval*	*paschal*	*preliminary*	*rental*
solid	*realistic*	*modern*	*protestant*	*same*	*reverse*
superior	*structured*	*moral*	*recessive*	*several*	*sliding*
sure	*varied*	*philosophical*	*sacred*	*top*	*thirteenth*

cluster 10 of 20		cluster 11 of 20		cluster 12 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
allergic	*carcinogenic*	*black*	*ceramic*	*bent*	*hysterical*
antibiotic	*coagulate*	*circular*	*cyclopean*	*bourgeois*	*inattentive*
artificial	*colorless*	*concave*	*fireproof*	*corrupt*	*irreligious*
dietary	*milky*	*crimson*	*legible*	*disreputable*	*lunatic*
fibrous	*nonfat*	*distinctive*	*rectilinear*	*domineering*	*opportunist*
liquid	*pulpy*	*fluorescent*	*sleek*	*evil*	*parochial*
mucous	*scented*	*incised*	*tucked*	*fascist*	*possessive*
powdery	*spongy*	*red*	*umber*	*jugular*	*resentful*
raw	*steamed*	*tubular*	*unglazed*	*pious*	*uncongenial*
synthetic	*vanilla*	*white*	*Venetian*	*warlike*	*unengaged*

cluster 13 of 20		cluster 14 of 20		cluster 15 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
bitter	*brokenhearted*	*affected*	*bottomed*	*abusive*	*appalling*
debilitating	*confused*	*buried*	*credited*	*deplorable*	*bias*
emotional	*delirious*	*distributed*	*jammed*	*exaggerated*	*capricious*
hopeless	*disturbed*	*given*	*owned*	*excessive*	*exorbitant*
odd	***odd***	*left*	*rose*	*illegal*	*hostile*
poor	*patchy*	*marked*	***scattered***	*simplistic*	*imprecise*
troubled	*regretful*	*modified*	*settled*	*undue*	*inelegant*
unhappy	*thirsty*	***scattered***	*shattered*	*unintentional*	*innocuous*
weird	***unhappy***	*used*	*surrounded*	*unproductive*	*unbalanced*
worst	*untidy*	*worn*	*sworn*	*wrong*	*unsound*

cluster 16 of 20		cluster 17 of 20		cluster 18 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
adrenal	*cesarean*	*American*	*Arabian*	***brilliant***	*adored*
alveolar	*endoscopic*	*Asian*	*Catalan*	*conspicuous*	*commanding*
bivariate	*hemorrhagic*	*Christian*	*Chinese*	*ecstatic*	*fantastic*
cardiovascular	*hyoid*	***Dutch***	*Cornish*	*extraordinary*	*favorite*
clinical	*intervertebral*	*English*	***Dutch***	*fitting*	*gallant*
diagnostic	*lobular*	*European*	*Haitian*	*great*	*halcyon*
neural	*monovalent*	*French*	*Hungarian*	*incomparable*	*loved*
peritoneal	*normotensive*	*Roman*	*Kurdish*	*perfect*	*superb*
spinal	*valved*	*Serbian*	*Taiwanese*	*singular*	*tragic*
ulcerative	*vesicular*	*Spanish*	*Thai*	*startling*	*undefeated*

cluster 19 of 20		cluster 20 of 20	
Frq-200	Rand-200	Frq-200	Rand-200
budgetary	*agrarian*	*aesthetic*	*clarion*
civil	*catechetical*	*artistic*	*contemporary*
criminal	*clandestine*	*classical*	*darkling*
marital	*constitutional*	*clever*	***dulcet***
mental	*curricular*	*colloquial*	*earthy*
national	*hourly*	*dreamy*	*falsetto*
nuclear	*intramural*	*hilarious*	*longhand*
parental	*qualitative*	*intimate*	*ponderous*
regional	*recreational*	*narrative*	*soothing*
social	*sectional*	*rhetorical*	*wry*

Appendix B

Why do the Curves go down?

As argued in Chapter 5, the predictive accuracy generally decreases across all models with time. At first glance, this seems counter-intuitive since predicting for later decades means the model has more observations to base its predictions off. However, as Figure B.1 shows, the average number of nouns to predict in each decade goes down with time. This general trend applies to both sets of true positives: only adjectives that first co-occur with a given noun n^* in decade $t + \Delta$, and also in any future decade. Consequently, precision scores generally decrease the chance of getting at least one prediction correct when the number of predictions is small. This likely explains why, as the average number of adjective predictions to make for each noun drops below 0.5 after the 1970s, the predictive accuracy of the models starts to drop rapidly. This trend is also reflected in the JSD curves (see Appendix C), as fewer predictions implies the empirical distribution becomes more peaked.

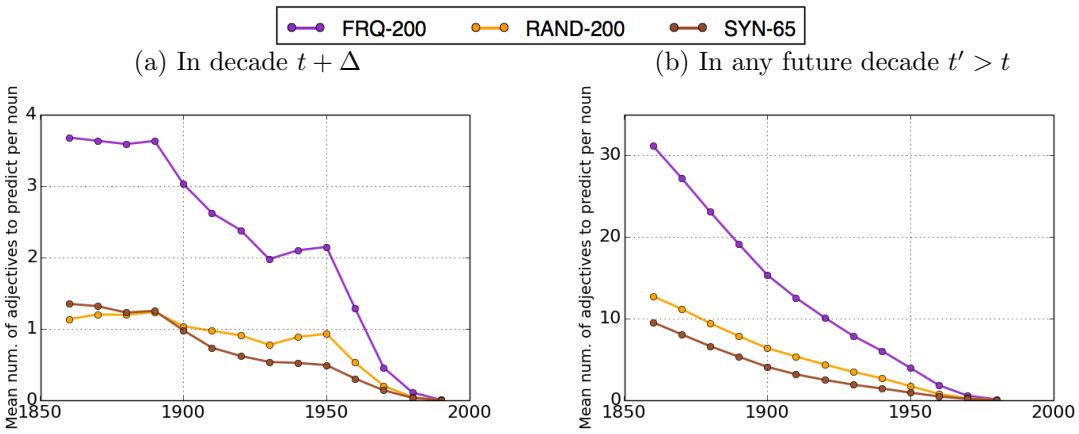


Figure B.1: The average number of novel adjective-noun pairs left for each model to predict across all times and adjective sets. This value is computed across all nouns for which a predictive model makes adjective predictions.

Appendix C

The Empirical Distribution

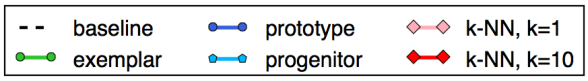
To further study model performance, we also examined how the model's predicted posterior distribution $p(a|n^*)^{(t+\Delta)}$ aligns with the empirical distribution of adjective-noun co-occurrences. We define the empirical posterior probability for that adjective a will co-occur with noun n^* as the token-based number of co-occurrences between a and n^* only if they first co-occur at time $t + \Delta$:

$$p_{\text{empirical}}\left(a|n^*\right)^{(t+\Delta)} \propto \text{count}(n^*, a, t+\Delta) \times \underbrace{\mathbb{1}\left[\text{count}(n^*, a, t+\Delta) \geq T\right]}_{\text{apply threshold to count}} \times \underbrace{\mathbb{1}\left[n^* \notin \{n\}_a^{(\leq t)}\right]}_{\text{disregard if previously attested}} .$$

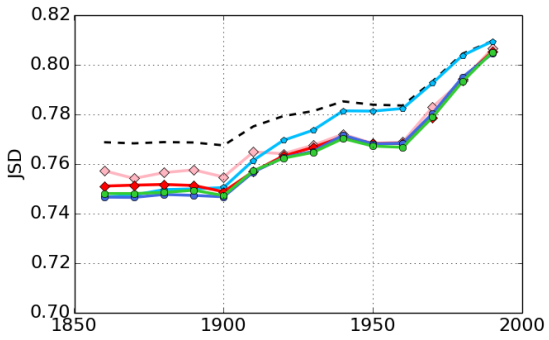
Here, $\text{count}(n^*, a, t + \Delta)$ is the (token-based) number of co-occurrences between a and n^* at time $t + \Delta$. In the above expression, we multiply the token-based count by $\mathbb{1}\left[\text{count}(n^*, a, t + \Delta) \geq 2\right]$ to threshold the number of co-occurrences at 2 (see Chapter 5.1 for more details about thresholding), and by $\mathbb{1}\left[n^* \notin \{n\}_a^{(\leq t)}\right]$ to ignore previously-attested adjective-noun pairs just as we do with the predicted posterior probability. We then train and evaluate our predictive models just as before, except we evaluate the expected Jensen-Shannon divergence (JSD) between the predicted and empirical posterior distributions over all nouns in $\mathcal{N}_{\mathcal{A}}^{(t)}$. We also learned kernel parameters for the exemplar and prototype models specific to this task by optimizing JSD.

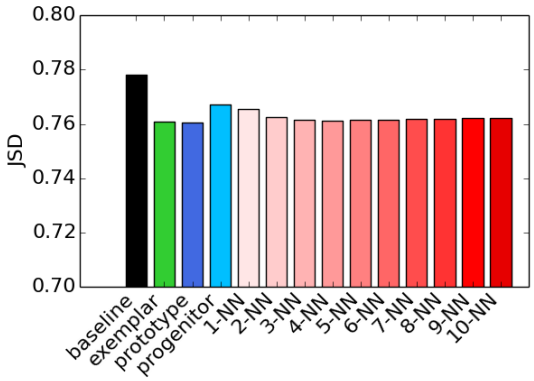
Figure C.1 reports the JSD between the predicted and empirical posterior distributions in each decade. In general, we see some noticeable differences in the rank-ordering of models here as compared with predictive accuracy. For instance, the 1-NN model obtains a lower JSD with the empirical posterior distribution much more frequently than the 10-NN model does on both the RAND-200 and SYN-65 adjective sets, even though the predictive accuracy suggests 10-NN better captures adjective extension. It's not clear why this occurs, especially since the rank-ordering agrees more with the results from

predictive accuracy on the Frq-200 adjective set.

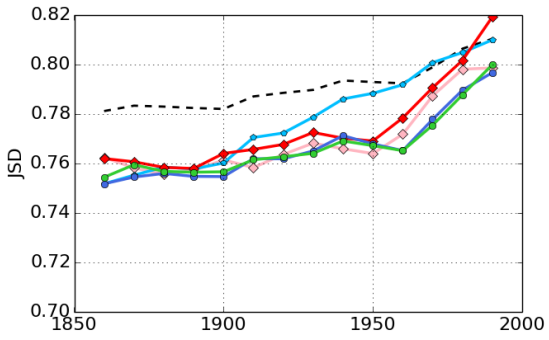


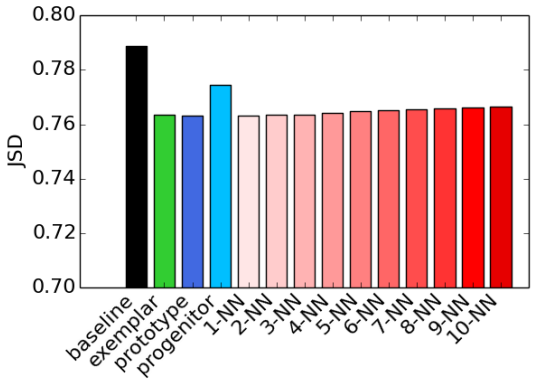
(a) Frq-200



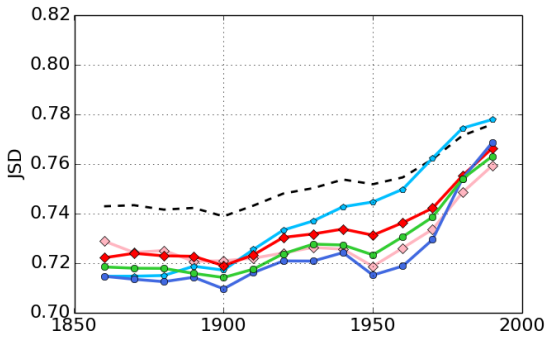


(b) Rand-200





(c) Syn-65



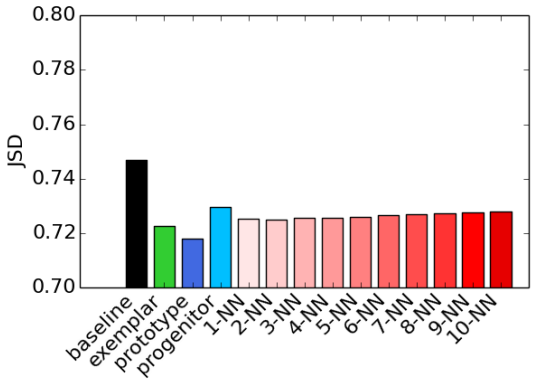


Figure C.1: Results from training and evaluating our predictive models with JSD across all three adjective sets. **Left:** The JSD between the predicted and empirical posterior distributions across all models and decades. **Right:** Aggregate JSD for all models (including k-NN from $k = 1$ to $k = 10$) across models and decades.

Appendix D

Learning via MAP Estimates

In Chapter 3.4, we described how to learn kernel parameters for each of our predictive models by maximizing precision. An alternative approach is to take a maximum a posteriori (MAP) estimate where we maximize the posterior probability of observing the attested adjective-noun pairs.

Let $(a_1, n_1), (a_2, n_2), \ldots, (a_L, n_L)$ be the novel adjective-noun pairings at time t. Given our predictive models, our MAP estimate of the kernel parameter $h^{(t)}$ follows

$$\begin{aligned}\widehat{h}^{(t)}_{\text{MAP}} &= \arg\max_h \prod_{i=1}^{L} p(a_i, h|n_i)^{(t)} \\ &= \arg\max_h \prod_{i=1}^{L} p(n_i|a_i, h)^{(t-\Delta)} p(a_i, h)^{(t-\Delta)}\end{aligned}$$

where the likelihood $p(n_i|a_i, h)^{(t-\Delta)}$ is computed via the exemplar or prototype likelihoods, and the prior $p(a_i, h)^{(t-\Delta)}$ is described in Chapter 3.2. Note that this prior is only dependent a_i and not h.

We tried this approach and evaluated our predictive models based on precision. As the results in Figure E.1 illustrate, a MAP estimate of kernel values is inferior to simply maximizing precision. The charts below report results for the exemplar and prototype models, and these results hold across all adjective sets. It's interesting how the drop in accuracy from precision- to MAP-based kernel parameters is affects the exemplar model more than it does the prototype model. As the next Appendix section shows, the kernel parameter values are also more similar in the prototype model than the exemplar model when comparing both methods for learning.

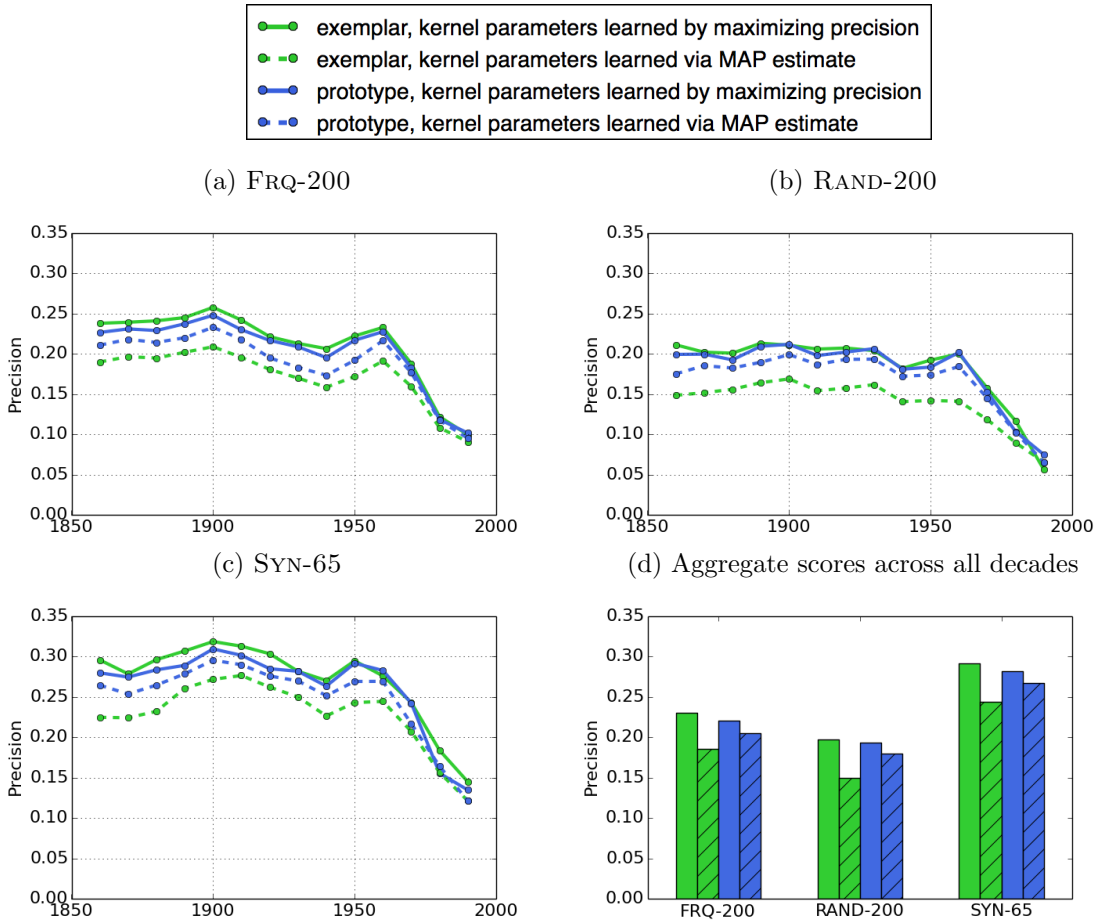


Figure D.1: A comparison of the exemplar and prototype models' predictive accuracy across all decades when the kernel parameter is learned via maximizing precision versus a MAP estimate on (a) Frq-200, (b) Rand-200, (c) Syn-65. Plot (d) gives the aggregate accuracy across all decades for each of the three adjective sets, within which the unstriped and striped bars correspond to kernel parameters learned through maximizing precision and a MAP estimate, respectively.

Appendix E

Kernel Parameters

We provide all kernel parameters that we used in our exemplar and prototype predictive models across all adjective sets and training objectives in the table below. Note that the left-hand column, t, gives the decade for which each kernel parameter was used to predict novel adjective-noun pairs. That is, the first row of the table where t = 1860s gives kernel parameters used to make predictions in the 1860s, but were learned based on adjective-noun pairs that emerged in the 1850s.

	FRQ-200					
	precision		JSD		MAP	
used to predict (t)	exemplar	prototype	exemplar	prototype	exemplar	prototype
1860s	0.1143	0.0500	0.0754	0.0308	0.5659	0.1266
1870s	0.1050	0.0589	0.0700	0.0295	0.5757	0.1191
1880s	0.1050	0.0600	0.0698	0.0293	0.5769	0.1331
1890s	0.1148	0.0648	0.0727	0.0300	0.6016	0.1468
1900s	0.1237	0.0650	0.0729	0.0293	0.6023	0.1340
1910s	0.1150	0.0672	0.0693	0.0301	0.6133	0.1306
1920s	0.1251	0.0576	0.0684	0.0296	0.6066	0.1563
1930s	0.1016	0.0596	0.0637	0.0283	0.6186	0.1729
1940s	0.1055	0.0550	0.0653	0.0286	0.6183	0.1586
1950s	0.0986	0.0539	0.0619	0.0300	0.6404	0.1593
1960s	0.1100	0.0621	0.0563	0.0291	0.6062	0.1113
1970s	0.1050	0.0621	0.0548	0.0283	0.6149	0.1189
1980s	0.1157	0.0625	0.0468	0.0260	0.6014	0.1024
1990s	0.0850	0.0850	0.0556	0.0254	0.6283	0.1020

	Rand-200					
	precision		JSD		MAP	
used to predict (t)	exemplar	prototype	exemplar	prototype	exemplar	prototype
1860s	0.1211	0.0570	0.0647	0.0300	0.5017	0.1147
1870s	0.1250	0.0500	0.0533	0.0291	0.5124	0.1043
1880s	0.1150	0.0550	0.0637	0.0278	0.5105	0.1055
1890s	0.1250	0.0549	0.0578	0.0285	0.5159	0.1106
1900s	0.1000	0.0543	0.0548	0.0286	0.5090	0.1070
1910s	0.1300	0.0512	0.0668	0.0301	0.5230	0.0963
1920s	0.1100	0.0602	0.0563	0.0279	0.5263	0.0965
1930s	0.1300	0.0563	0.0500	0.0273	0.5382	0.0876
1940s	0.1188	0.0500	0.0445	0.0278	0.5273	0.0877
1950s	0.1050	0.0609	0.0483	0.0278	0.5134	0.0869
1960s	0.1050	0.0500	0.0485	0.0271	0.5271	0.0908
1970s	0.1050	0.0450	0.0457	0.0241	0.5309	0.0777
1980s	0.0700	0.0550	0.0360	0.0243	0.5464	0.0775
1990s	0.1300	0.0250	0.0452	0.0207	0.5214	0.0801
	Syn-65					
	precision		JSD		MAP	
used to predict (t)	exemplar	prototype	exemplar	prototype	exemplar	prototype
1860s	0.1154	0.0463	0.0736	0.0251	0.6169	0.1172
1870s	0.1250	0.0650	0.0684	0.0251	0.6385	0.1236
1880s	0.1050	0.0450	0.0569	0.0223	0.6703	0.1484
1890s	0.1033	0.0504	0.0640	0.0224	0.6852	0.1498
1900s	0.1026	0.0650	0.0583	0.0212	0.6984	0.1711
1910s	0.1450	0.0850	0.0507	0.0205	0.7168	0.1824
1920s	0.1414	0.0950	0.0581	0.0200	0.7779	0.1747
1930s	0.1641	0.0742	0.0354	0.0181	0.7650	0.1620
1940s	0.1500	0.0850	0.0419	0.0144	0.7529	0.1561
1950s	0.1250	0.0600	0.0410	0.0150	0.7507	0.1597
1960s	0.1550	0.0600	0.0104	0.0125	0.7013	0.1056
1970s	0.1203	0.0500	0.0249	0.0102	0.7989	0.1313
1980s	0.0800	0.0350	0.0356	0.0103	0.8177	0.1340
1990s	0.1000	0.0350	0.0509	0.0104	1.2439	0.6566

The kernel parameters listed in the above tables are also illustrated in the following plots.

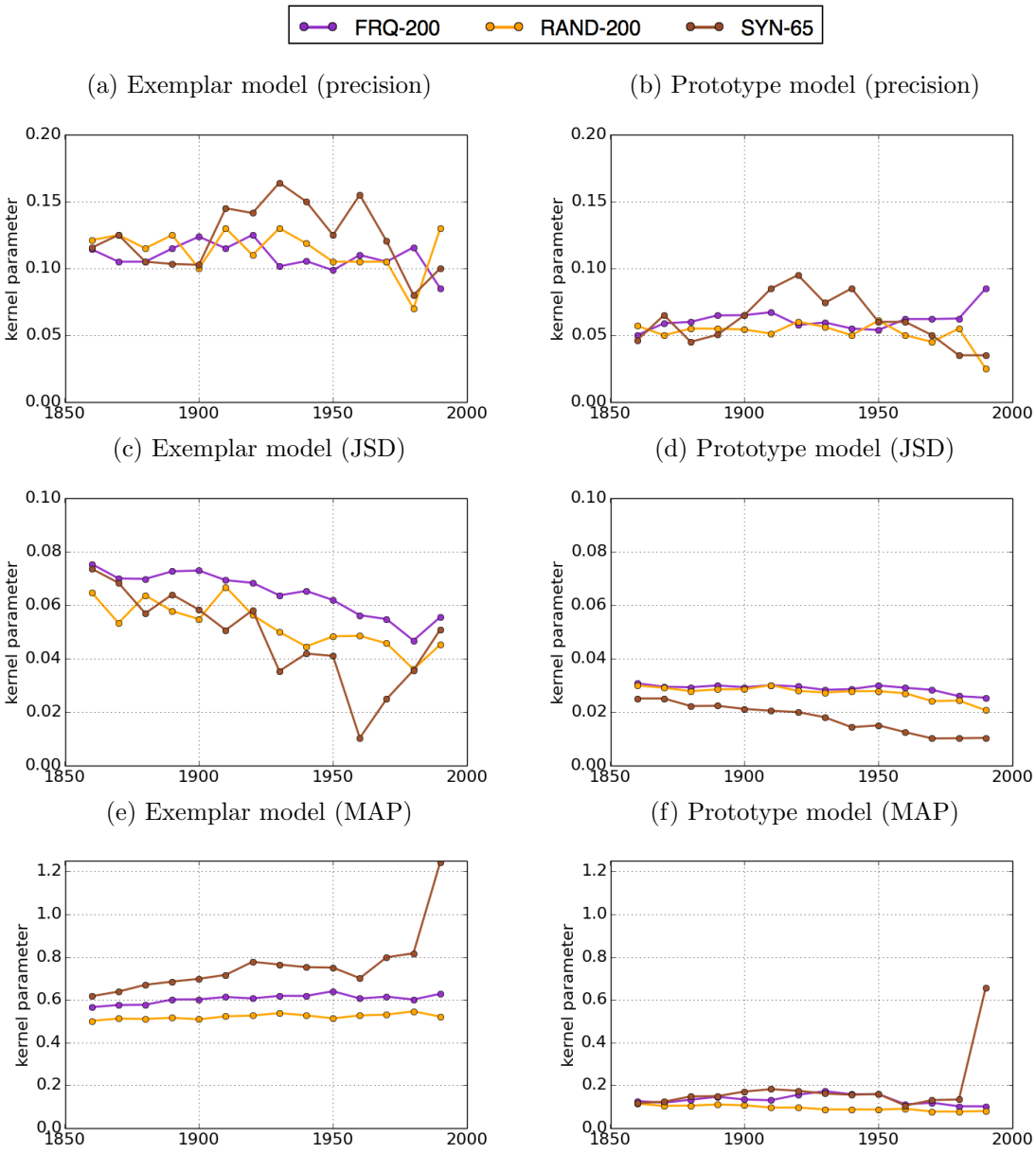


Figure E.1: The kernel parameters learned by the exemplar (left) and prototype (right) models with precision-based, and JSD-based, and MAP objectives across all adjective sets as a function of time.