To appear in Cognition.

Chaining and the growth of linguistic categories

Amir Ahmad Habibi[a], Charles Kemp[b], Yang Xu[c,*]

[a]*Department of Computing Science, University of Alberta*
[b]*School of Psychological Sciences, University of Melbourne*
[c]*Department of Computer Science, Cognitive Science Program, University of Toronto*

Abstract

We explore how linguistic categories extend over time as novel items are assigned to existing categories. As a case study we consider how Chinese numeral classifiers were extended to emerging nouns over the past half century. Numeral classifiers are common in East and Southeast Asian languages, and are prominent in the cognitive linguistics literature as examples of radial categories. Each member of a radial category is linked to a central prototype, and this view of categorization therefore contrasts with exemplar-based accounts that deny the existence of category prototypes. We explore these competing views by evaluating computational models of category growth that draw on existing psychological models of categorization. We find that an exemplar-based approach closely related to the Generalized Context Model provides the best account of our data. Our work suggests that numeral classifiers and other categories previously described as radial categories may be better understood as exemplar-based categories, and thereby strengthens the connection between cognitive linguistics and psychological models of categorization.

Keywords: Category growth; Semantic chaining; Radical categories; Categorization; Exemplar theory; Numeral classifiers

[*] Corresponding author
Email address: yangxu@cs.toronto.edu (Yang Xu)

1. Introduction

Language users routinely face the challenge of categorizing novel items. Over the past few decades, items such as *emojis*, *blogs* and *drones* have entered our lives and we have found ways to talk about them. Sometimes we create new categories for novel items, but in many cases we assign them to existing categories. Here we present a computational analysis of the cognitive process by which categories extend in meaning over time.

Lakoff and other scholars [1, 2, 3, 4] have suggested that linguistic categories grow over time through chaining, a process that links novel items to existing items that are semantically similar, hence forming chain-like structures of meaning [1]. Although Lakoff briefly suggests how chaining applies to semantic categories (e.g. the concept of "climbing'), his two most prominent examples of chaining involve grammatical categories. The first example is the classifier system of Dyirbal (an Australian Aboriginal language), which groups together nouns that may not seem closely related on the surface. For instance, the word *balan* may precede nouns related to women, fire and dangerous things. The second example is the Japanese classifier *hon*, which can be applied to a variety of long thin objects such as pencils, sticks and trees. Where an English speaker might say "one pencil," a Japanese speaker must insert the appropriate classifier (here *hon*) between the numeral and the noun. Although *hon* is most typically applied to long thin objects, it can also be applied to martial arts contests using swords (which are long thin objects), and to medical injections (which are carried out using long, thin needles). Martial arts contests and medical injections have little in common, but both can be connected to central members of the *hon* category through a process of chaining.

In Lakoff's work the notion of chaining is coupled with the notion of centrality, which proposes that a category is organized around a central core. Combining chaining with centrality leads to the notion of a *radial category*, or one that can be characterized as a network of chains that radiate out from a center [1, 5]. Subsequent work in cognitive linguistics relaxes the idea of a single center and allows that radial categories may have "several centers of comparable importance" (Palmer & Woodman, 2000, p 230), but is still committed to the idea that some members of a radial category are privileged by virtue of their centrality. In principle, however, the notions of chaining and centrality can be decoupled. Consider, for example, a category that is constructed by starting with one element and repeatedly adding a new element that is similar to a randomly chosen member of the category. This generative process seems consistent with the notion of chaining, but the categories it produces may take the form of sprawling networks rather than collections of chains radiating out from a center.

Many discussions of chaining within cognitive linguistics are heavily influenced by Rosch and her prototype theory of categorization (e.g., Geeraerts, 1997), but this literature has been largely separate from the psychological literature on computational models of categorization [7, 8]. The modeling literature includes many comparisons between exemplar models and prototype models of categorization, and the question of whether categories have a central core lies at the heart of the difference between the two approaches. Exemplar models proposes that the representation of a category is no more than an enumeration of all members of the category, but prototype models propose that category representations incorporate some additional element such as a prototype, a central tendency or a set of core examples.[1] Decoupling chaining from centrality means that the process of chaining is potentially compatible with both prototype-based and exemplar-based accounts of categorization, and opens up the possibility of formal accounts of chaining that build on exemplar models like the Generalized Context Model (GCM, Nosofsky, 1986) that have achieved notable success as psychological models of categorization. Here we evaluate a suite of formal models, including a prototype model and a family of exemplar models, and find that an exemplar model closely related to the GCM provides the best account of category growth over time. Our results are broadly consistent with previous work on computational models of categorization, which often finds that exemplar theory outperforms prototype theory when instances of the two are put to the test.

Following Lakoff we focus on grammatical categories, and as a case study we

[1] In her later work Rosch explicitly suggested that "prototypes do not constitute a theory of representation for categories" (Rosch, 1978, p 40). Much of the literature on prototype theory, however, does make representational claims.

consider how Chinese numeral classifiers have been applied to novel nouns over the past fifty years. As for Japanese classifiers, Chinese classifiers are obligatory when a noun is paired with a numeral, e.g., *one [classifier$_x$] person* or *two [classifier$_y$] documents*. Although we focus on Chinese classifiers, numeral classifiers are found in many other languages around the world, and have been extensively studied by cognitive psychologists, linguists, and anthropologists [11, 1, 12, 13, 14, 1]. For instance, Allan (1977) has suggested that classifiers across languages often capture perceptual properties such as shape and size, and Aikhenvald (2000) has suggested that classifiers also capture more abstract features such as animacy. Although previous scholars have explored how people assign classifiers to nouns [15, 16], most of this work has not been computational. Our approach goes beyond the small amount of existing computational work [17, 18, 19, 20, 21] by analyzing historical data and focusing on the application of classifiers to novel nouns.

There are at least three reasons why numeral classifiers provide a natural venue for testing computational theories of category extension. First, they connect with classic examples such as Lakoff's analysis of *hon* that are central to the cognitive linguistics literature on chaining and category extension. Second, classifiers are applied to nouns, which form a broad and constantly-expanding part of the lexicon, and therefore offer many opportunities to explore how linguistic categories are applied to novel items. Third, the item classified by a term like *hon* is typically the noun phrase that directly follows the classifier, which makes it relatively simple to extract category members from a historical corpus (e.g., via part-of-speech tags).

Our work goes beyond Lakoff's treatment of classifiers in three important ways. First, we present a computational framework that allows us to evaluate precise hypotheses about the mechanism responsible for chaining. Second, we test these hypotheses broadly by analyzing a large set of classifiers and their usage in natural contexts, instead of considering a handful of isolated examples. Third, as mentioned already our space of models includes exemplar-based approaches that have not been explored in depth by previous computational accounts of chaining. Previous scholars have given exemplar-based accounts of several aspects of language including phonetics, phonology, morphology, word senses, and constructions [22, 23, 24, 8, 25, 26], and our approach builds on and contributes to this tradition.

Our approach also builds on recent computational work that explores formal models of chaining in the historical emergence of word meanings. In particular, Ramiro et al (2018) demonstrated that neighbourhood-based chaining algorithms can recapitulate the emerging order of word senses recorded in the history of English. This work found that the best-performing algorithm was a nearest-neighbour model that extends the semantic range of a word by connecting closely related senses. Two earlier studies report that the same nearest-neighbour model also accounts for container naming across languages [28, 29]. This paper compares a suite of models including the nearest-neighbour model that was successful in previous work. We find that our historical data on the growth of Chinese classifiers is best explained by a model that adjusts the nearest-neighbour approach in several ways that are consistent with the GCM [10], an influential exemplar-based model of categorization. Our results therefore suggest that the same categorization mechanisms studied in lab-based tests of the GCM may help to explain how real-world linguistic categories extend over time.

2. Theoretical framework

Figure 1 illustrates how semantic chaining might influence which Chinese classifier is applied to a novel noun. We begin by assuming that nouns correspond to points in a semantic space. Given a novel noun, the classifier for that noun can then be predicted based on classifiers previously applied to nearby nouns in the space. In Figure 1 the novel noun is referendum, which entered the Chinese lexicon around the year 2000. Nearby nouns in the space have two different classifiers: 次 (cì) is used for nouns like "employment," "funding" and "speech" (shown as orange circles) and 项 (xiàng) is used for nouns like "extension" and "estimate" (shown as blue triangles). The year in which each noun emerged has been estimated from a corpus described later, and in this corpus the first appearance of "referendum" happens to be paired with cì (the orange classifier).

The notion of chaining suggests that "referendum" is classified by linking it with one or more previously encountered nouns that are similar in meaning. In Figure 1, “referendum” has been linked with 11 nearby nouns. According to the corpus, the nouns closest to “referendum” tend to be paired with cì, which may explain why cì is also used for “referendum.” Iterating this process through time leads to chaining because the classification of “referendum“ influences classifications of subsequently encountered nouns – in particular, assigning cì to “referendum“ means that the same classifier is more likely to be used for novel nouns near “referendum.”

The informal characterization of chaining just presented leaves many details unspecified, and the following sections attempt to fill in some of these gaps. The next section presents a formal framework for modelling category growth over time. We then specify a set of competing hypotheses about the function that determines how the classifications of nearby nouns influence the classification of a novel noun. A subsequent section discusses the nature and origin of the semantic space that captures similarity relationships between the nouns.

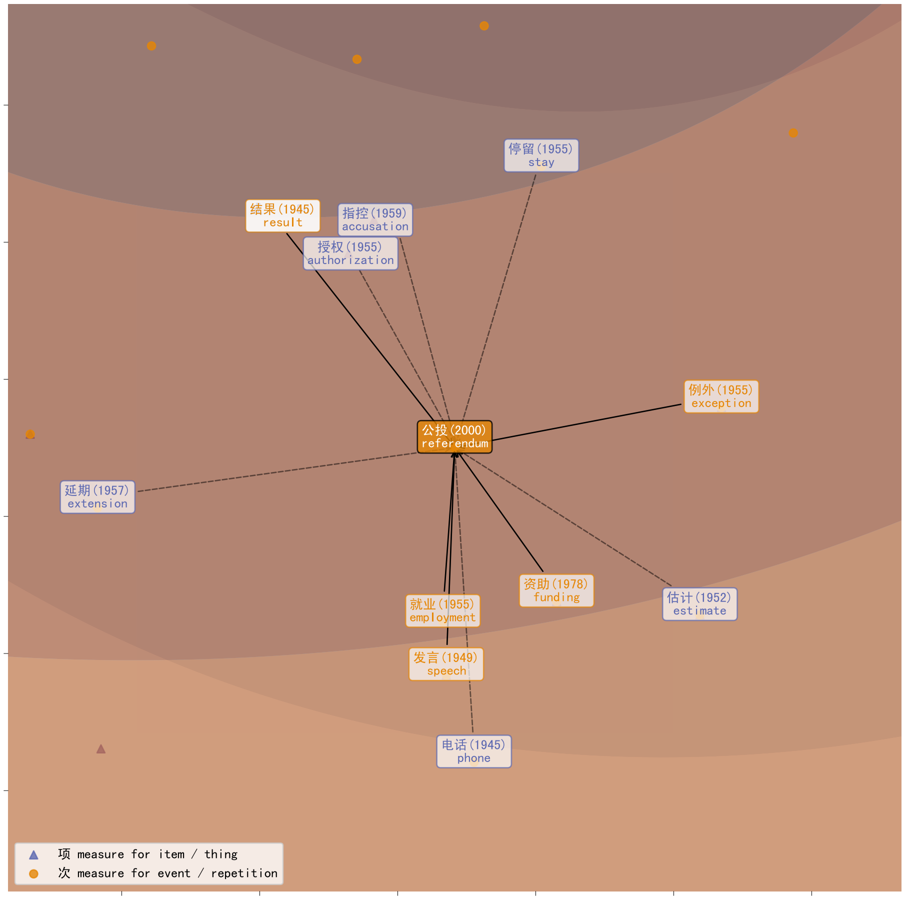

Figure 1: An illustration of chaining in Chinese classifiers. "Referendum" entered the language around 2000, and nearby nouns in semantic space are shown as orange circles or blue triangles depending on which of two classifiers our corpus pairs them with. The closest nouns belong to the orange category, and "referendum" is also assigned to this category. For visual clarity only selected nouns in the space have been labeled. The background colors indicate how strongly each classifier is favored in each region. The blue category is favored in the darker regions near the top, and the orange category is favored elsewhere in the space.

2.1. A probabilistic formulation of category extension

Let c denote an existing category (e.g. a classifier), x denote an item (e.g. a noun), and t and $t+$ denote current and future time respectively. We formulate category extension as the problem of assigning a novel item x^* to an existing category:

$$p(c|x^*)_{t+} \propto f(x^*|c)_t \times p(c)_t \qquad (1)$$

Equation 1 casts category extension as sequential probabilistic inference, where the goal is to predict future category labels at time $t+$ given the likelihood $f(x^*|c)_t$ and prior $p(c)_t$ at the current time t. This formulation postulates that the probability of assigning x^* to category c is jointly influenced by the probability of seeing that item given the category, and the prior probability of choosing that category.

The general framework in Equation 1 can be used to explore how categories from many parts of speech (including nouns, adjectives, verbs, and adverbs) are applied to novel items. Here we focus on classifiers, and therefore assume that category c is a classifier and that item x^* is a noun.

2.2. *Formal hypotheses about chaining*

To formally characterize the chaining mechanism we must specify the likelihood function in Equation 1, which captures how novel nouns relate to existing nouns, and the prior distribution in Equation 1, which captures the *a priori* probability of each classifier. We will evaluate a set of models that make different assumptions about these two components.

2.2.1. *Likelihood function $f(x^*|c)_t$*

We considered thirteen likelihood functions that specify how nouns x previously paired with a classifier c (i.e., $x \in c$) might influence whether the classifier is applied to a novel noun x^*. Representative examples of these likelihood functions are illustrated in Figure 2. Each function assumes that classifier choice depends on similarity relationships between the novel noun and familiar nouns, where similarity is defined by exponentiating the negative Euclidean distance $d(\cdot,\cdot)$ between nouns in semantic space [30, 30, 31, 10]:

$$sim(n_1, n_2) = \exp(-d(n_1, n_2)^2) \tag{2}$$

Most previous computational models of chaining [28, 29, 27] rely on a nearest-neighbour (1nn) approach that assigns a novel item to the same category as the nearest familiar item. Let n_c^k denote the number of items with category label c among the k

items most similar to x. 1nn can then be formulated using the function

$$f(x^*|c) = \begin{cases} 1 & \text{if } n_c^1 = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

1nn corresponds exactly to previous computational work on chaining [28, 29, 27], but we suspected that the 1-neighbor assumption might be too strict. We therefore evaluated a set of k-nearest-neighbor classifiers that assign a category label to x that matches the most common label among the k items most similar to x:

$$f(x^*|c) = \begin{cases} 1 & \text{if } n_c^k = \max\limits_{c' \in C} n_{c'}^k \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where C is the set of all categories and $\max\limits_{c' \in C} n_{c'}^k$ is the frequency of the most common category among the k items most similar to x. We evaluated a total of 10 different models (including 1nn) that set k to all integers between 1 and 10 inclusive.

Although some exemplar-based models rely on a nearest neighbor approach, the dominant exemplar-based approach considers relationships between a novel item and *all* previously encountered items, weighting each one by its similarity to the novel item:

$$f(x^*|c) = \sum_{x \in c} sim(x^*, x) = \sum_{x \in c} \exp(-d(x^*, x)^2) \tag{5}$$

In the psychological literature the most influential exemplar-based model is the Generalized Context Model (GCM) [10]. The GCM can be formulated as a Bayesian model [32] that relies on the exemplar approach and includes a sensitivity parameter s that controls the rate at which similarity decreases with distance in semantic space:

$$f(x^*|c) = \sum_{x \in c} \exp(-sd(x^*, x)^2) \tag{6}$$

Large values of s mean that similarity falls off rapidly with distance, which in turn means that only the nearest exemplars to a novel item influence how it is classified. Smaller values of s lead to broader regions of influence. We will refer to the likelihood function in Equation 6 as the *exemplar* approach, and the function in Equation 5 as the *exemplar (s=1)* approach.

All likelihood functions introduced so far are broadly compatible with the exemplar-based view of categories. As mentioned earlier, however, many cognitive linguists view chaining as a mechanism for generating radial categories, and the notion of a radial category is derived from Rosch's prototype theory. Ideally we would like to evaluate a prototype model with a likelihood function that captures Lakoff's views about radial categories, and in particular his view of classifier categories like Japanese "hon." To our knowledge such a model has never been formulated, but the psychological literature does include simple prototype models of categorization. Here we evaluate one such model which assumes that the prototype of a category is the average of all exemplar types that belong to the category [33].

$$\text{prototype}_c = \frac{1}{|c|} \sum_{x \in c} x \tag{7}$$

$$f(x^*|c) = sim(x^*, \text{prototype}_c) = \exp(-d(x^*, \text{prototype}_c)^2) \tag{8}$$

This approach allows the prototype of a category to change over time as new exemplars are added to the category, and postulates that category extension occurs by linking a novel item to the prototype that is closest in semantic space. Even if a novel item lies closer to the prototype of category A than that of category B, the handful of exemplars closest to the item may belong to category B, which means that the prototype and exemplar models sometimes make different predictions. Although the prototype model evaluated here is useful as a starting point, developing and evaluating more sophisticated computational models of prototype theory is an important challenge for future work, and we return to this issue in the general discussion.

Although the thirteen likelihood functions capture different assumptions about chaining, they are comparable in model complexity. The only parameter tuned in our model comparison is s, the sensitivity parameter used by the exemplar model. To avoid giving this model an unfair privilege we set this parameter based on held-out data. In principle one could consider sensitivity-weighted versions of the other likelihood functions, but for our purposes these variants turn out to be equivalent to the versions without sensitivity weights. We will evaluate our models based on the proportion of correct classifications that they predict, and adding sensitivity weights to the nearest-neighbour and prototype models changes the confidence associated with their classifications but not the classifications themselves.

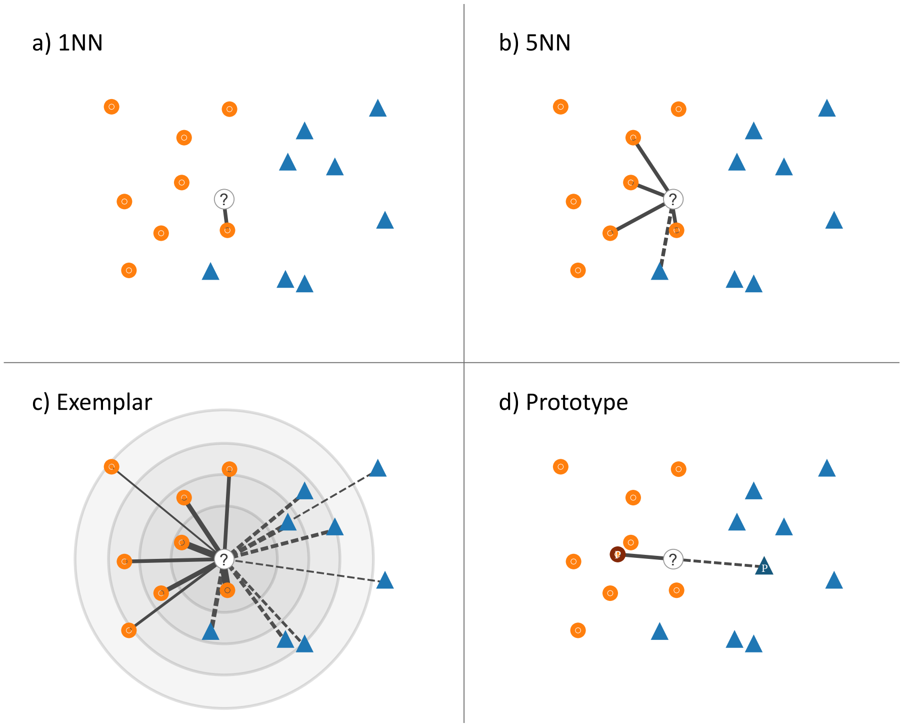

Figure 2: Illustrations of four likelihood functions. Each panel assumes that there are two categories shown as circles and triangles, and that a novel item shown as a question mark must be assigned to one of these categories. Edges show relationships between the novel item (question mark) and previously encountered item. In (c), the edges differ in thickness because items closer to the novel item are weighted more heavily. In (d), the two nodes labelled "P" are prototypes of the two categories.

2.2.2. Prior distributions

We used two prior distributions over classifiers. The first is uniform, and the second is a size-based prior that assumes that the prior probability of a classifier is proportional to the number of different nouns a classifier was paired with previously, i.e., type frequency. The size-based prior connects with the idea that categories or words tend to attract new items according to a rich-get-richer process [34, 35, 36].

Combining the two priors with the thirteen likelihoods produces a set of 10 possible models. For comparison with these models we considered two additional baselines. The first is a random guess model that assigns each novel item to a classifier chosen uniformly at random. The second is a max-category model that assigns a novel item at time $t+$ to the classifier with maximum type frequency up to time t (i.e. the classifier that has been paired with the greatest number of different nouns). These baselines can be interpreted as models that use either a uniform or a size-based prior but assume that the likelihood function in Equation 1 is constant.

3. Exemplar vs Prototype models: Simulated data

Although the exemplar and prototype models are formally different, it is possible that they lead to categories with similar statistical properties. For example, even though an exemplar-based category includes no central core, it is still possible that categories grown according to the exemplar model tend to end up roughly convex in shape with members arranged around a central region. To examine whether and how the exemplar and prototype models produce different kinds of categories, we compared these models using simulated data.

Simulation procedure.

We simulate category growth in a continuous two-dimensional space bounded by [0,1] along each dimension. Each run begins with three randomly chosen points that serve as seed exemplars for three categories. We then generate additional random points, one at a time, and record the category labels assigned to each point by the exemplar and prototype models. In addition to the prototype model described above, we also consider a static prototype model where the category prototypes are fixed throughout time and correspond to the three seed exemplars. Figure 3 illustrates one simulation run and shows category growth according to the three models over 100 iterations. Although all three models are given the same sequence of points, they produce different category systems by the end of the run. We used two quantitative measures to compare systems produced by the models: category size and category discriminability.

Expected category size.

The first measure quantifies the average size of categories generated by each models. The prototype models are consistent with the notion of radial categories, and we expected that they would tend to produce compact categories with members arranged around a central prototype. The exemplar model, however, allows more scope for categories that consist of elongated chains or other arbitrary shapes.

We measured category size as the area of the convex hull that includes all members of a category. Expected category size is then computed as the average of this quantity across the three categories in the simulation. Figure 3 shows that expected category size is greater for the exemplar model than for the two prototype models, supporting the intuition that exemplar-based categories tend to be less compact than radial categories. Figure 4 (left panel) confirms this finding across 500 simulated runs. We found that the exemplar model generally produces an expected category size that is substantially greater than the prototype model with a moving core, and both of these models generate categories that are larger on average than those produced by the static prototype model.

Category discriminability.

The second measure quantifies the degree to which categories are discriminable (or separable) under each model. High discriminability means that there are relatively few ambiguous cases near the boundary between two categories, and near-optimal systems of categories will tend to have high discriminability. If exemplar-based categories tend to be elongated, one might expect that they intertwine in complex ways and are therefore less discriminable than the more convex categories produced by the prototype models.

We quantify category discriminability using an extension of Fisher's linear discriminant that allows for more than two categories. Given $k = 3$ categories with category means m_1, m_2, m_3 and covariances $\Sigma_1, \Sigma_2, \Sigma_3$, we compute Fisher's discriminant ratio r by weighing the cross-category separability (of the means) against the pooled within-category variabilities (based on the covariance determinants):

$$r = \frac{d(m_1,m_2)+d(m_1,m_3)+d(m_2,m_3)}{|\Sigma_1|+|\Sigma_2|+|\Sigma_3|} \tag{9}$$

Here $d()$ represents Euclidean distance. A high discriminability value indicates that the categories are highly separable, and is achieved, for example, if inter-category distances are high and within-category variability is low.

Figure 3 shows that the exemplar and prototype models both produce categories with equally high discriminability, and that both models produce more discriminable categories than the static prototype model. Even though exemplar-based categories are less compact than prototype-based categories, Figure 3 suggests that this difference in compactness has no implications for discriminability, which is consistent with previous findings from container naming that neighbourhood-based chaining tends to yield categories that are near-optimally structured for efficient communication [29].

Taken together, our simulations suggest two general messages. First, the fact that exemplar and prototype models produce category systems with similar levels of discriminability suggests that the two models lead to outcomes that are similar in key respects. As a result, careful analyses may be needed to distinguish between these two competing models of category growth. Second, the results for category size reveal that exemplar and prototype models do lead to patterns of category growth with statistically different properties. This finding means that analyses of real-world categories (e.g. Chinese classifiers) can plausibly aim to determine whether the process underlying the growth of these categories is closer to an exemplar model or a prototype model.

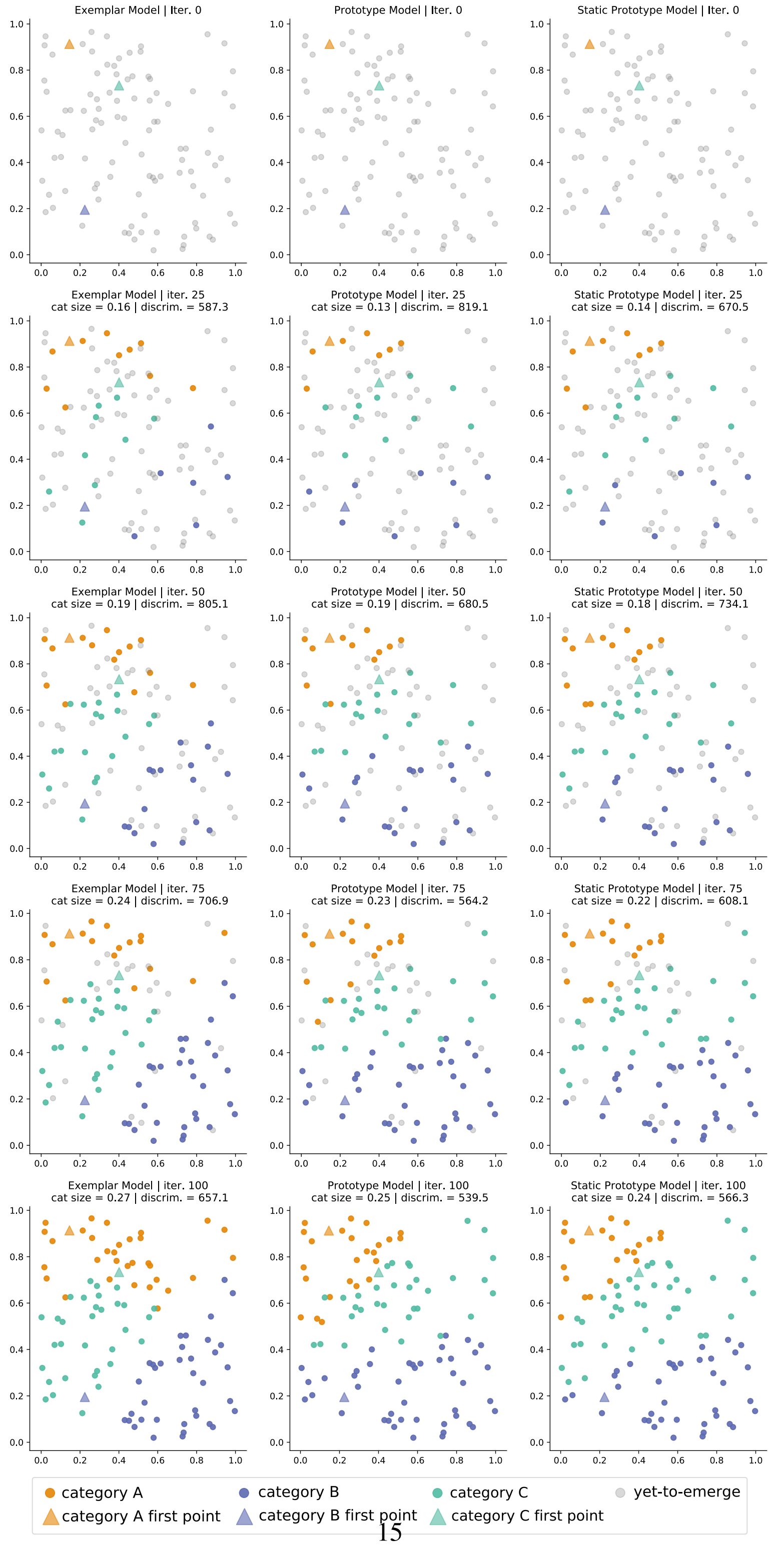

Figure 3: Simulated category growth under the exemplar and prototype models. A static version of the prototype model is also considered where the prototype remains fixed (as opposed to dynamic) over time.

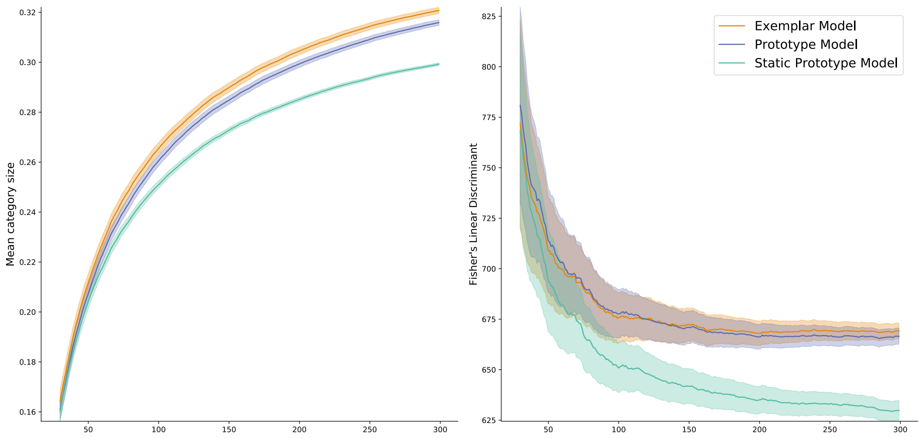

Figure 4: Category compactness and discriminability analysis of the exemplar and prototype models. Category size (left panel) and Fisher discriminant ratio (right panel) is calculated under each model over multiple simulation runs with random initial points. Shaded areas correspond to 95% confidence bands.

4. Analysis of Chinese classifiers through time

We next applied the models to the growth of Chinese classifiers through time. Doing so required three primary sources of data: a large repository of web-scraped Chinese (classifier, noun) pairs ; 2) historical time stamps that record the first attested usage of each (classifier, noun) pair; and 3) a semantic space capturing similarity relationships between nouns.

4.1. Classifier data

We worked with a comprehensive list of (classifier, noun) pairs compiled by Morgado da Costa, Bond, & Gao (2016), who kindly made their data available. This resource includes a total of 57966 classifier-noun pairs that incorporate 185 classifiers and 32414 unique nouns. To reduce noise we removed 28473 classifier-noun pairs that appeared only once according to the frequency tags. We also removed classifiers that were paired with fewer than 10 distinct nouns and are therefore highly specific.

4.2. Historical time stamps

To time stamp the (classifier-noun) pairs, we used N-gram data from the Google Books corpus of simplified Chinese [37], which tracks the frequency of each word pair over the period 1940-2003. We specifically searched for (classifier,noun) pairs that had a "_NOUN" tag for their noun part.

4.3. Semantic space

The likelihood term in Equation 1 requires some representation of semantic relations among nouns. Although several choices are available for specifying a semantic space, including dictionary definitions [27] and human similarity judgments, we chose to focus on word embeddings derived from large-scale natural language use.

We used pre-trained word2vec embeddings of English translations of the Chinese nouns. Chinese word embeddings are also available, but English embeddings are preferred for our purposes because the Chinese embeddings are based on a contemporary corpus that includes the same noun-classifier pairings that our models are aiming to predict. [2]

To establish mappings between nouns in Chinese and word embeddings in English, we used Google Translate to identify English equivalents of the Chinese nouns. Some Chinese nouns are mapped to English phrases, and we included only cases where a Chinese noun could be paired with a single English word. [3] We worked with 4045 unique nouns that were available both in the English word2vec model through translation and in the Google N-grams, yielding 8371 total pairs of classifier-noun usages for our analyses.

4.4. Model evaluation

Each of the models described previously was evaluated based on its ability to predict classifiers assigned to novel nouns over the period 1951 to 2003. We assessed these predictions incrementally over time: for each historical year where a novel classifier-noun usage appeared according to the time stamps, we compared the held-out true classifier with the model-predicted classifier that had the highest posterior probability

[2] We thank an anonymous reviewer for pointing out that Chinese embeddings smuggle in information about noun-classifier pairings.

[3] Three native speakers of Mandarin Chinese independently inspected a sample of 100 Chinese-English noun pairs and considered 98, 97, and 95 of those translations to be acceptable, respectively.

(i.e., term on the left of Equation 1) given the novel noun. In cases where a noun appeared with multiple classifiers, we excluded classifiers that had previously appeared with the noun when computing model predictions (i.e., we only make predictions about classifers that are paired with a noun for the first time). This procedure ensures that there are no repeated predictions from any of the models.

To estimate the sensitivity parameter s of the exemplar models, for each year, we used data from the years before (i.e. data from 1941 until that year) and performed an optimization within the range of 0.1 to 100 to identify the s that maximized the performance of the model for the nouns that emerged during the previous year. [4] Appendix A includes the estimated values of the sensitivity parameter for these models.

4.5. Results

Figure 5 summarizes the overall predictive accuracies of the models. The best performing model overall was the exemplar model ($s = 1$) with size-based prior. All models are based on types rather than tokens: for example, $P(c)$ is proportional to the number of types that classifier c is paired with rather than the combined count of tokens of these types. Appendix B includes results for token-based models and shows that they perform uniformly worse than their type-based equivalents. We return to this finding in the general discussion, but focus here on the results for type-based models, and begin by considering the contribution made by each component of these models.

Contribution of the prior.

The baseline model with size-based prior and constant likelihood achieved an accuracy of 29.6%, which is substantially better than random choice (accuracy of 1.6% among 127 classifiers). Figure 5 shows that the size-based prior led to better performance than the uniform prior. In 12 out of 13 cases, a size-based model performed better than the same model with uniform prior ($p < 0.002, n = 13, k = 12$ under binomial test). [5]

[4] A line search was performed with a step size 0.1 in the range 0.1–1.0 and a step size of 1.0 in the range 1.0–100.0.

[5] In all k-nearest-neighbor models, we used the size-based prior when there is a tie among the classifier categories when they share the same number of nearest neighbors to a noun. In the uniform-prior case, we randomly choose a classifier if there is a tie.

Our results therefore support the idea that being paired with many types of nouns in the past makes a classifier especially likely to be applied to novel nouns.

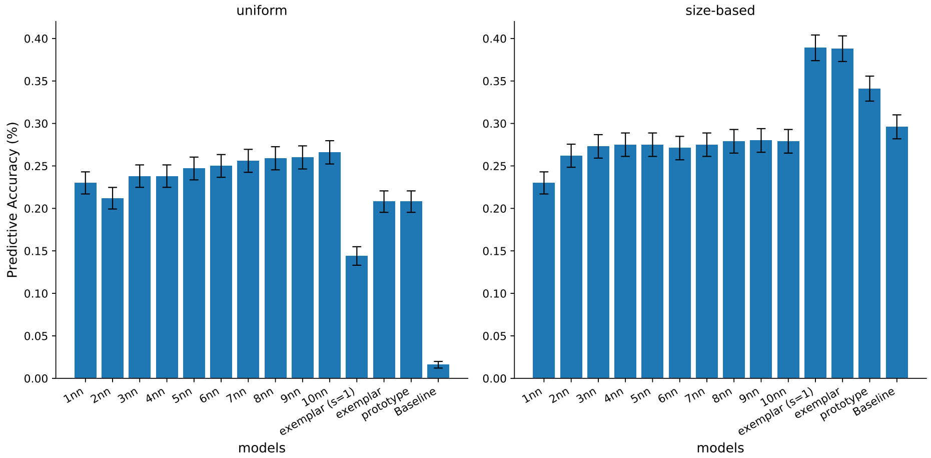

Figure 5: Summary of predictive accuracies achieved by all models under the two priors.

Contribution of the likelihood function.

Figure 5 (right panel) shows that the prototype likelihood function leads to worse performance than the exemplar model ($s = 1$). Recall, however, that the prototype models evaluated here are extremely simple, and that more sophisticated formulations of prototype theory could well perform better.

The results show that optimization of the sensitivity parameter (s) did not improve performance of the exemplar model. This parameter is not a free parameter but was rather optimized based on held-out data, and the fact that this optimization did not improve on the default setting $s = 1$ suggests that the optimization process probably led to overfitting. Alternative methods for optimizing s may be able to improve on the default setting, but for our purposes the key result is that there is a clear gap between the exemplar model and both the prototype and nearest neighbour models.

Previous formal approaches to chaining have often used a 1NN approach [27, 29], but our results suggest that considering more than one neighbor of a novel item may be beneficial in the case of classifier extension. None of the k-nearest-neighbor models performed better than the exemplar model (or the prototype model), but the incremental performance from 1 neighbor to high-order neighbors suggests that approximation of neighborhood density matters in the process of chaining. The exemplar model can be considered as a soft but more comprehensive version of the *k*-nearest-neighbor model class, where all semantic neighbors are considered and weighted by distance to the novel item in prediction.

Figure 6 confirms our findings by showing the time courses of predictive accuracy for the models, highlighting three aspects: 1) models with the size-based prior generally achieved better performance than models with a uniform prior; 2) the best overall exemplar model ($s = 1$) with the size-based prior is consistently superior to the other competing models (including the prototype model) through the time period of investigation; 3) increasing the order of nearest neighbors improves model prediction. Our results therefore support a key theoretical commitment of the GCM, which proposes that categorization judgments are made by computing a weighted sum over all previous category exemplars.

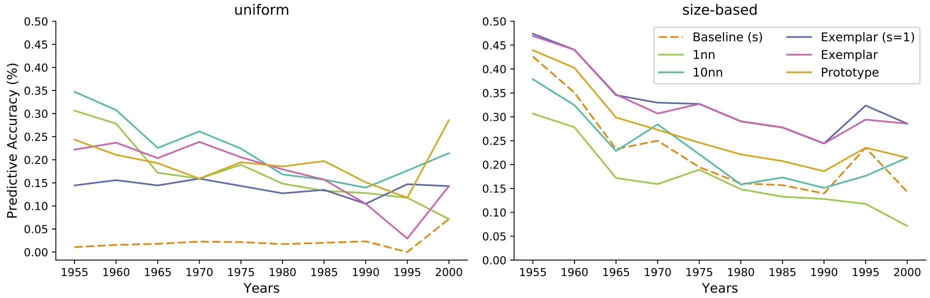

Figure 6: Predictive accuracies of representative models in the use of 127 Chinese classifiers at 5 year intervals between 1955 and 2000.

Results for individual classifiers.

Table 1 shows examples of classifiers paired with nouns that emerged relatively recently along with predictions of the best model (the exemplar model $s = 1$). For example, the noun 网民 (netizen) emerged in 2001 according to our data, and the model successfully predicts that 名 (classifier for people) is the appropriate classifier. The model makes sensible predictions of classifier usage even when it was considered incorrect. For instance, 并购 (merger) was paired with 宗 (classifier for events involving transactions or business related things) in our data, and the model predicts 起 , which is a classifier used for describing events.

Figure 7 shows precision and recall for individual classifiers based on the same model. The model achieves high precision for a classifier if it is mostly correct when it chooses that classifier. For example, 尊 is typically applied to sculpture, and the model is always correct when it chooses this classifier. High recall is achieved for a classifier if the model chooses that classifier in most cases in which it is actually correct. The recall for 尊 is low, suggesting that the model fails to apply this classifier in many cases in which it is correct.

The classifier with highest recall is 个, which is a generic classifier that is extremely common. Recall tends to be greatest for the most frequent classifiers, which is expected given that the model uses a frequency-based prior. The classifiers with highest precision are specialized classifiers that are relatively rare, which means that they are rarely chosen by the model in cases where they do not apply.

Table 1: Examples of novel nouns, English translations, ground-truth Chinese classifiers and predictions of the exemplar model ($s = 1$) with size-based prior.

Noun	English translation	Year of emergence	True classifier	Model-predicted classifier
博客	blog	2003	个	个
网民	netizen	2001	名	名
公投	referendum	2000	次	次
帖子	(Internet) post	1999	篇	名
股权	equity	1998	批	项
并购	merger	1998	宗	起
网友	(Internet) user	1998	名	个
机型	(aircraft) model	1997	款	位
玩家	player	1997	名	名

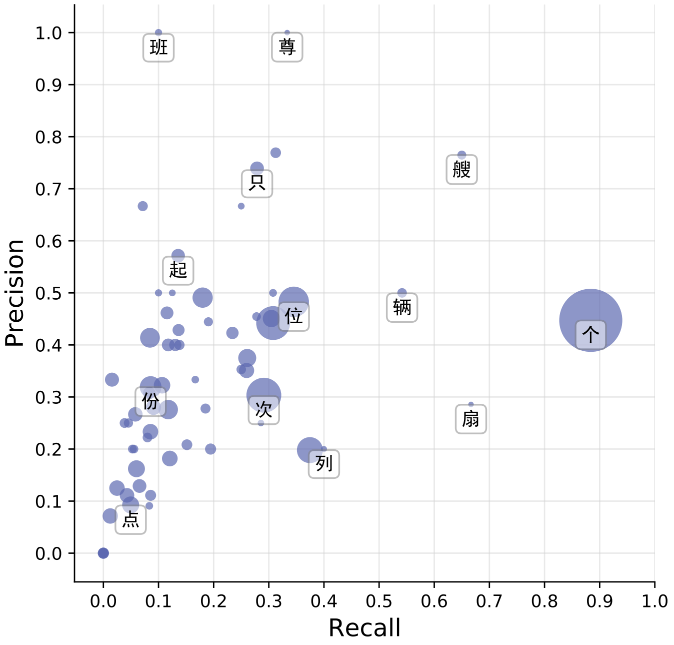

Figure 7: Precision and recall of individual classifiers based on the best exemplar model. Marker size is proportional to category size (i.e., number of different nouns paired with a classifier).

5. Discussion

We presented a principled computational account of the historical extension of linguistic categories. Our approach is based on a probabilistic framework that allowed us to formulate and test a large space of models that make predictions about how Chinese classifiers were extended to novel nouns over the past half century. The results suggest that classifications of novel nouns are influenced by classifier frequency and by classifications of previous nouns that are similar in meaning. As suggested earlier, our approach connects with and extends prior work in cognitive linguistics and categorization, and we now elaborate on some of these connections.

5.1. Connections with cognitive linguistics

Classifiers are prominently discussed by Lakoff and others as examples of categories that grow through a process of chaining, but most work in this area focuses on qualitative analyses of a handful of examples. Our work builds on previous treatments by considering a set of computational models of chaining and evaluating them across a relatively large historical corpus.

To our knowledge, our work is the first to apply a computational model of chaining to the domain of numeral classifiers, but previous papers have used formal models of chaining to study container naming [28, 29] and word senses in a historical dictionary [27]. Each of these contributions evaluates several formal models including a weighted exemplar approach and finds that a nearest-neighbour approach performs best. In contrast, we found that a weighted exemplar approach closely related to the GCM provided the best account of our data. The reasons for these different conclusions are not entirely clear. As suggested earlier, the weighted exemplar approach reduces to a nearest-neighbour approach when the sensitivity parameter s becomes large, which means that the weighted exemplar approach should always perform at least as well as the nearest-neighbour approach for some value of s. For generic values of s, it seems possible that the nature of the semantic representation influences the performance of the weighted exemplar approach. Previous models of chaining used semantic representations based on human similarity judgments [28, 29] and a taxonomy constructed by lexicographers [27], and it is possible that the word embeddings used in our work are especially well suited to a weighted exemplar approach.

The literature on cognitive linguistics suggests some directions in which our work can be extended. Lakoff presents chaining as a mechanism that leads to radial categories, which are organized around one or more central cores and therefore qualify as prototype categories. We evaluated a simple prototype model drawn from the psychological literature, and found that this model performed worse than an exemplar-based approach. This result provides some initial support for the idea that numeral classifiers are best understood as exemplar-based categories, but definitive support would require the evaluation of more sophisticated prototype models that better capture the way in which linguists think about radial categories. A key challenge is to develop semantic representations that better capture the full richness of word meanings (e.g., multi-modal representations that combine linguistic and extra-linguistic cues such as visual and conceptual relations). For example, consider Lakoff's proposal that Japanese *hon* is extended from long thin objects to medical injections because injections are given using long, thin needles. Our work represents nouns as points in a semantic space, and although useful for some purposes this representation does not adequately capture the way in which multiple concepts (e.g. the purpose of an injection, the setting in which it might occur, and the instrument by which it is administered) come together to create a richly-textured meaning. Developing improved semantic representations is a major research challenge, but one possible approach is to combine the word embeddings used in our work with more structured representations [38] that identify specific semantic relations (e.g. agent, patient, instrument) between concepts.

A second important direction is to extend our approach to accommodate mechanisms other than chaining that lead to meaning extension over time. For example, metaphor (e.g., *grasp*: "physical action"→"understanding") has been proposed as a key cognitive force in semantic change [39], and recent work provides large-scale empirical evidence of this force operating throughout the historical development of English [40]. An apparent difference between chaining and metaphor is that chaining operates within localized neighborhoods of semantic space, but metaphoric extension may link items that are relatively remote (as in the case of *grasp*). Metaphorical extension (e.g., *mouse*: "rodent"→"computer device") could also rely on perceptual information that is beyond the scope of our current investigation. As suggested already, a richer representation of semantic space will be needed, and it is possible that the chaining mechanisms proposed here will capture some aspects of metaphorical extension when operating over that richer representational space.

5.2. Connections with the categorization literature

Our work is grounded in the psychological literature on categorization, and joins a number of previous projects [41, 42, 43] in demonstrating how computational models can be taken out of the laboratory and used to study real-world categories. Our best performing model is a weighted-exemplar approach that is closely related to the GCM and that goes beyond nearest-neighbor models in two main respects. First, it classifies a novel item by comparing it to many previously-observed exemplars, not just a handful of maximally-similar exemplars. Second, it uses a prior that favors classifiers that have previously been applied to many different items. Both ideas are consistent with the GCM, and our results suggest that both are needed in order to account for our data as well as possible.

Our best model, however, differs from the GCM in at least one important respect. Throughout we focused on type frequency rather than token frequency. For example, the size-based prior in our models reflects the number of types a classifier was previously paired with, not the number of previous tokens of the classifier. Models like the GCM can be defined over types or tokens [44], but it is more common and probably more natural to work with tokens rather than types. The empirical evidence from the psychological literature on type versus token frequencies is mixed: some studies find an influence of type frequency [45], but others suggest that token-based models perform better than type-based models [44, 46]. It seems likely that type frequencies and token frequencies both matter, but predicting how the two interact in any given situation is not always straightforward.

Our finding that the exemplar model performed better given type frequencies rather than token frequencies is broadly compatible with an extensive linguistic literature on the link between type frequency and the productivity of a construction [47, 48, 49, 50]. For example, consider two past-tense constructions that both include a slot for a verb. If the two constructions occur equally often in a corpus (i.e. token frequency is equal) but one construction occurs with more different verbs (i.e. has higher type frequency) than the other, then the construction with higher type frequency is more likely to be extended to a novel verb. The link between type frequency and productivity is supported by both corpus analyses and modeling work. For example, our results parallel the work of Albright & Hayes (2003), who present a model of morphological learning that achieves better performance given type frequencies instead of token frequencies.

Although the link between type frequency and productivity has been clearly established, token frequency also affects linguistic generalizations. For instance, Bybee (1985) suggests that high token frequency is negatively related to productivity, because a construction that appears especially frequently with one particular item may be learned as an unanalyzed whole instead of treated as a structure with slots that can be filled by a range of different items. Items with high token frequencies may also be treated as category prototypes [53, 49], which means that token frequency will be relevant when developing prototype models more sophisticated than the one evaluated here. Previous theories [47, 54, 48, 55] and computational models [56, 57] of language learning have incorporated both type frequency and token frequency, and extending our approach in this direction is a natural goal for future work.

The psychological literature suggests at least two additional directions that future work might aim to pursue. We considered how an entire speech community handles new items that emerge over a time scale of years or decades, but psychological models often aim to capture how individuals learn on a trial-by-trial basis. Accounting for the classifications made by individual speakers is likely to require ideas that go beyond our current framework. For example, individuals might be especially likely to reuse classifiers that have recently occurred in a conversation, and there may be kinds of selective attention that operate at a timescale of seconds or minutes and are not well captured by the models used in our work. Psychologists have studied how numeral classifiers are applied in the lab [15, 16], and there is an opportunity to combine this experimental approach with the modeling approach that we have developed. A second important direction is to explore how children acquire numeral classifiers over the course of development. If applied to a corpus of child-directed speech, our model could potentially make predictions about errors made by children as they gradually learn the adult system of numeral classifiers.

6. Conclusion

We presented a framework for exploring how linguistic categories change over time. We took numeral classifiers as a case study, and evaluated the claim that these categories grow through a process of chaining. Our results support this claim but suggest that the underlying mechanism is more like a weighted exemplar model than the nearest-neighbor approach advocated by previous work on chaining. Although numeral classifiers are often described as radial categories, our results provide some initial evidence that the growth of these categories may be better captured by exemplar theory than by prototype theory.

Although we focused on numeral classifiers, our approach is relatively general, and could be used to explore how other linguistic categories change over time. In recent years historical corpora have become more accessible than ever before, and we hope that future work can build on our approach to further explore how linguistic change is shaped by cognitive processes of learning and categorization.

7. Acknowledgements

We thank Luis Morgado da Costa for sharing the classifier dataset, and three anonymous reviewers for helpful comments on the manuscript. This work is supported by an NSERC Discovery Grant, a SSHRC Insight Grant, and a Connaught New Researcher Award to YX.

Appendix A. Estimated values of the sensitivity parameter

Figures A.8 and A.9 show the estimated values of the sensitivity parameter for the exemplar models under different choices of prior and semantic space, based on types and tokens separately.

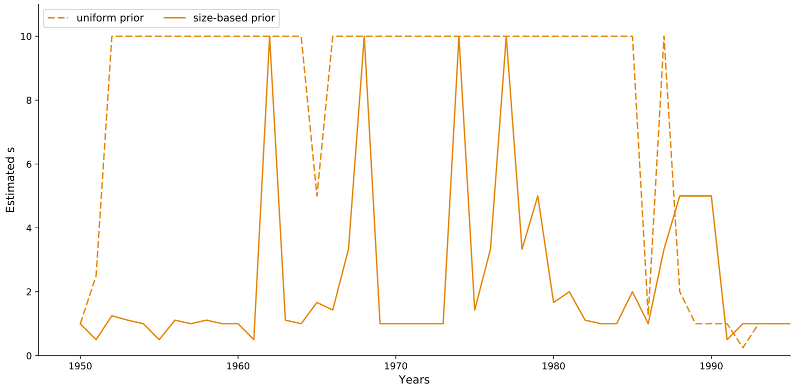

Figure A.8: Estimated optimal values of the sensitivity parameter (s) from the type-based models.

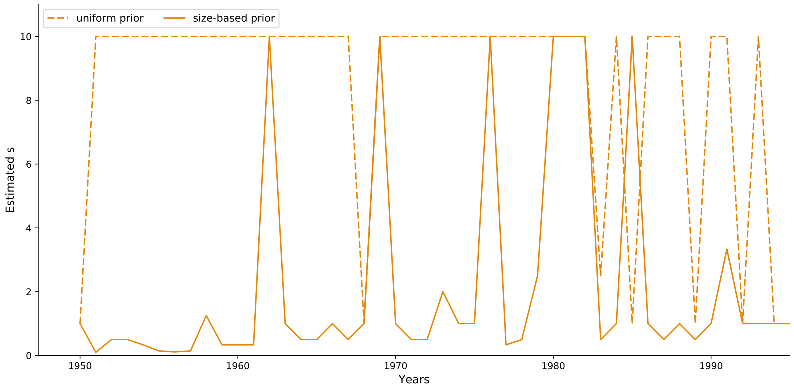

Figure A.9: Estimated optimal values of the sensitivity parameter (s) from the token-based models.

Appendix B. Token-based models

The models in the main text are based on types rather than tokens, and Figure B.10 shows corresponding results for token based exemplar and prototype models (keeping *k*-nearest-neighbor models the same because token-based results for low-order *k*'s are 520 effectively invariant and similar to a type-based 1nn model). For the prototype model, we defined the prototype of a category as the frequency-weighted average:

$$\text{prototype}_c = E[x|c] = \sum_{x \in c} x\, p(x|c) = \sum_{x \in c} x \frac{\text{freq}(x)}{\sum_{x' \in c} \text{freq}(x')} \tag{B.1}$$

.

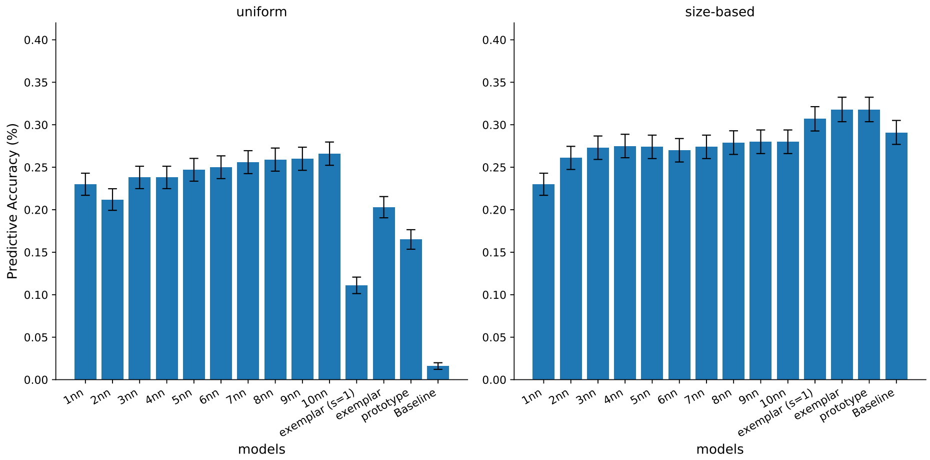

Figure B.10: Summary of predictive accuracies achieved by all token-based models under the two priors.

Appendix C. Supplementary material

Code and data used for our analyses are available on GitHub at `https://github.com/AmirAhmadHabibi/ChainingClassifiers`. Pre-trained English word2vec embeddings are available at `https://code.google.com/archive/p/word2vec/` [58, 59, 60], and the N-gram data we used from the Google Book corpus are available at `http://storage.googleapis.com/books/ngrams/books/datasetsv2.html` [61].

References

[1] G. Lakoff, Women, fire, and dangerous things: What Categories Reveal About The Mind, University of Chicago Press, Chicago, 1987.

[2] J. L. Bybee, R. D. Perkins, W. Pagliuca, The evolution of grammar: Tense, aspect, and modality in the languages of the world, University of Chicago Press, Chicago, 1994.

[3] D. Geeraerts, Diachronic prototype semantics: A contribution to historical lexicology, Oxford University Press, Oxford, 1997.

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

[5] B. Lewandowska-Tomaszczyk, Polysemy, prototypes, and radial categories, in: The Oxford handbook of cognitive linguistics, 2007.

[6] G. B. Palmer, C. Woodman, Ontological classifiers as polycentric categories, as seen in Shona class 3 nouns, in: M. Pütz, M. Verspoor (Eds.), Explorations in Linguistic Relativity, Amsterdam; Philadelphia; J. Benjamins Pub. Co; 1999, 2000, pp. 225–248.

[7] F. Polzenhagen, X. Xia, Language, culture, and prototypicality, in: The Routledge Handbook of Language and Culture, Routledge, 2014, pp. 269–285.

[8] S. Chandler, The analogical modeling of linguistic categories, Language and Cognition 9 (1) (2017) 52–87.

[9] E. Rosch, Principles of categorization, in: E. Rosch, B. B. Lloyd (Eds.), Cognition and categorization, Lawrence Erlbaum Associates, New York, 1978, pp. 27–48.

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

[11] A. Y. Aikhenvald, Classifiers: A typology of noun categorization devices, Oxford University Press, Oxford, 2000.

[12] K. Allan, Classifiers, Language 53 (1977) 285–311.

[13] R. Dixon, R. M. W. Dixon, R. M. Dixon, The Dyirbal language of North Queensland, Cambridge University Press, Cambridge, 1972.

[14] B. Berlin, A. K. Romney, Descriptive semantics of Tzeltal numeral classifiers, American Anthropologist 66 (1964) 79–98.

[15] M. Y. Gao, B. C. Malt, Mental representation and cognitive consequences of Chinese individual classifiers, Language and Cognitive Processes 24 (2009) 1124–1179.

[16] J. H. Y. Tai, Chinese classifier systems and human categorization, in: Honor of William S.-Y. Wang: Interdisciplinary studies on language and language change, Pyramid Press, Taipei, 1994, pp. 479–494.

[17] H. Guo, H. Zhong, Chinese classifier assignment using SVMs, in: Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, 2005.

[18] N. Peinelt, M. Liakata, S. Hsieh, ClassifierGuesser: A context-based classifier prediction system for chinese language learners, in: Proceedings of the 8th International Joint Conference on Natural Language Processing, 2017.

[19] H. M. S. Wen, H. E. Gao, F. Bond, Using WordNet to predict numeral classifiers in Chinese and Japanese, in: Proceedings of the 6th Global WordNet Conference, 2012.

[20] L. Morgado da Costa, F. Bond, H. Gao, Mapping and generating classifiers using an open Chinese ontology, in: Proceedings of the 8th Global WordNet Conference, 2016.

[21] M. Zhan, R. Levy, Comparing theories of speaker choice using a model of classifier production in mandarin chinese, in: Proceedings of the 17th Annual Conference

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, p. 1997–2005.

[22] R. Skousen, Analogical modeling of language, Springer Science & Business Media, 1989.

[23] J. B. Pierrehumbert, Exemplar dynamics: Word frequency, lenition and contrast, Typological Studies in Language 45 (2001) 137–158.

[24] E. Keuleers, Memory-based learning of inflectional morphology, Ph.D. thesis, Universiteit Antwerpen (2008).

[25] J. L. Bybee, Usage-based theory and exemplar representations of constructions, in: T. Hoffmann, G. Trousdale (Eds.), The Oxford handbook of construction grammar, Oxford University Press, Oxford, 2013.

[26] R. Ramsey, An exemplar-theoretic account of word senses, Ph.D. thesis, Northumbria University (2017).

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

[28] S. A. Sloman, B. C. Malt, A. Fridman, Categorization versus similarity: The case of container names, Oxford University Press, New York, 2001, pp. 73–86.

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

[30] R. N. Shepard, Stimulus and response generalization: tests of a model relating generalization to distance in psychological space., Journal of Experimental Psychology 55 (1958) 509?523.

[31] R. M. Nosofsky, Luce's choice model and Thurstone's categorical judgment model compared: Kornbrot's data revisited, Attention, Perception, & Psychophysics 37 (1985) 89–91.

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

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

[34] J. R. Anderson, The adaptive nature of human categorization, Psychological Review 98 (1991) 409–429.

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

[36] Y. Luo, Y. Xu, Stability in the temporal dynamics of word meanings, in: Proceedings of the 40th Annual Meeting of the Cognitive Science Society, 2018.

[37] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al., Quantitative analysis of culture using millions of digitized books, Science 331 (2011) 176–182.

[38] C. F. Baker, C. J. Fillmore, J. B. Lowe, The Berkeley Framenet project, in: Proceedings of the 17th International Conference on Computational linguistics–Volume 1, Association for Computational Linguistics, 1998, pp. 86–90.

[39] E. Sweetser, From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure, Cambridge University Press, Cambridge, 1991.

[40] Y. Xu, B. C. Malt, M. Srinivasan, Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium, Cognitive Psychology 96 (2017) 41–53.

[41] G. Storms, P. De Boeck, W. Ruts, Prototype and exemplar-based information in natural language categories, Journal of Memory and Language 42 (1) (2000) 51–73.

[42] W. Voorspoels, W. Vanpaemel, G. Storms, Exemplars and prototypes in natural language concepts: A typicality-based evaluation, Psychonomic Bulletin & Review 15 (3) (2008) 630–637.

[43] R. M. Nosofsky, C. A. Sanders, M. A. McDaniel, A formal psychological model of classification applied to natural-science category learning, Current Directions in Psychological Science 27 (2) (2018) 129–135.

[44] R. M. Nosofsky, Similarity, frequency, and category representations., Journal of Experimental Psychology: Learning, Memory, and Cognition 14 (1) (1988) 54.

[45] A. Perfors, K. Ransom, D. Navarro, People ignore token frequency when deciding how widely to generalize, in: Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 36, 2014.

[46] L. W. Barsalou, J. Huttenlocher, K. Lamberts, Basing categorization on individuals and events, Cognitive Psychology 36 (1998) 203–272.

[47] J. Bybee, Language, usage and cognition, Cambridge University Press, Cambridge, 2010.

[48] J. Barðdal, Productivity: Evidence from case and argument structure in Icelandic, John Benjamins, Amsterdam, 2008.

[49] A. E. Goldberg, Constructions at work: The nature of generalization in language, Oxford University Press, Oxford, 2006.

[50] T. C. Clausner, W. Croft, Productivity and schematicity in metaphors, Cognitive science 21 (3) (1997) 247–282.

[51] A. Albright, B. Hayes, Rules vs. analogy in english past tenses: A computational/experimental study, Cognition 90 (2) (2003) 119–161.

[52] J. L. Bybee, Morphology: A study of the relation between meaning and form, John Benjamins, Amsterdam/Philadelphia, 1985.

[53] J. Bybee, D. Eddington, A usage-based approach to Spanish verbs of ‘becoming’, Language 82 323–355.

[54] D. V. Wilson, Gradient conventionalization of the Spanish expression of ‘becoming’quedar (se)+ ADJ in seven centuries, John Benjamins, Amsterdam/Philadelphia, 2018, pp. 175–198.

[55] J. Barðdal, The semantic and lexical range of the ditransitive construction in the history of (North) Germanic, Functions of Language 14 (1) (2007) 9–30.

[56] M. Johnson, T. L. Griffiths, S. Goldwater, Adaptor grammars: A framework for specifying compositional nonparametric bayesian models, in: Advances in Neural Information Processing Systems, 2007, pp. 641–648.

[57] T. J. O’Donnell, Productivity and reuse in language: A theory of linguistic computation and storage, MIT Press, 2015.

[58] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Workshop at ICLR, 2013.

[59] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: NIPS, 2013.

[60] T. Mikolov, W.-T. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: NAACL-HLT, 2013.

[61] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, , J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, E. L. Aiden, Quantitative analysis of culture using millions of digitized books, Science 331 (6014) (2011) 176–182. doi:10.1126/science.1199644.