J. Appl. Prob. **55,** 390–401 (2018)
*doi:*10.1017/jpr.2018.25

EFFECTS OF LIMITING MEMORY CAPACITY ON THE BEHAVIOUR OF EXEMPLAR DYNAMICS

B. GOODMAN * ** AND
P. F. TUPPER,* *** *Simon Fraser University*

Abstract

Exemplar models are a popular class of models used to describe language change. Here we study how limiting the memory capacity of an individual in these models affects the system's behaviour. In particular, we demonstrate the effect this change has on the extinction of categories. Previous work in exemplar dynamics has not addressed this question. In order to investigate this, we will inspect a simplified exemplar model. We will prove for the simplified model that all the sound categories but one will always become extinct, whether memory storage is limited or not. However, computer simulations show that changing the number of stored memories alters how fast categories become extinct.

Keywords: Exemplar model; linguistics; language change; extinction

2010 Mathematics Subject Classification: Primary 91F20
Secondary 70F99

1. Introduction

In spoken and written language, there are instances where there are two or more variants of a word, each of which is equivalent from the point of view of communication. We can think of instances of the word as belonging to one of two or more categories. For example, a population might pronounce the word 'either' as both 'ee-ther' and 'eye-ther'. Another example is when there are different spellings of a word. In Figure 1, which was generated by Google Books™ Ngram Viewer [7], we present a comparison between the usage of the word 'cider' and its archaic spelling 'cyder'. In the year 1800, 'cyder' seems to have been the more popular spelling but it has become practically extinct since then. As we see in this example, it is possible for a category to become extinct, passing out of usage.

Here we study a model for just this kind of category extinction. One popular class of models used to research the evolution of spoken and written language are exemplar models, first introduced by Nosofsky [8], [9]. Nosofsky hypothesized that people store detailed memories of stimuli they are exposed to which are called exemplars [15]. Johnson [5] showed that exemplar theory could be applied to model speech perception.

Exemplar theory models language use in one individual. Exemplars are detailed memories of utterances of sounds, each with its own category label. Categories are formed of all exemplars with a given category label. Exemplars are represented as vectors where each dimension represents a phonetic variable such as fundamental frequency or tongue height. Each exemplar will have a weight (or activation) associated with it, representing how predominant or recent

Received 5 April 2017; revision received 11 December 2017.
* Postal address: Department of Mathematics, Simon Fraser University, 8888 University Dr., Burnaby, BC, V5A 1S6, Canada.
** Email address: b_goodman@live.ca
*** Email address: pft3@sfu.ca

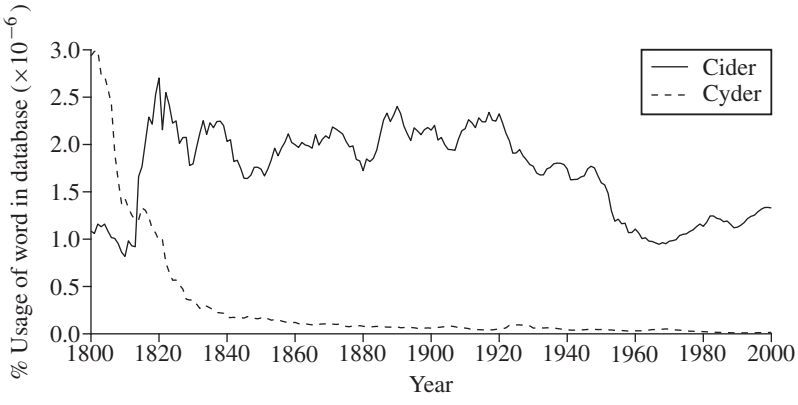

FIGURE 1: Comparison of the usage of 'cider' (*solid*) and its archaic spelling 'cyder' (*dashed*) within a corpus of books between the years 1800 and 2000. The y-axis represents the percentage of the usage of the words in the entire database. This image was generated by Google Ngram Viewer [7].

the memory of the sound is. In many exemplar models, these weights decay exponentially over time [10].

Exemplar dynamics builds on exemplar theory by creating a production-perception loop between two individuals with their own stored exemplars. Exemplar dynamics was first used by Pierrehumbert [10] to model speech production and perception. Following this, many linguists have used exemplar dynamics to model spoken and written language; see, for example, [3], [4], [15], [16], and [17].

In exemplar dynamics, there are usually two individuals speaking to one another. Each individual has a store of labelled exemplars. At every time-step a new sound is produced by a speaker, which is then perceived by the listener and classified based on the listener's stored exemplars. The way in which the sound is produced varies depending on the model. Usually a new sound is produced randomly by adding noise and bias to a preexisting exemplar. The listener usually categorizes sounds based on their 'closeness' to the cloud of exemplars stored for each category. The weights of the exemplars decay at each time-step, and the process is repeated. Newly categorized sounds become a part of the perception process, continually evolving the system [10].

The extinction of a category occurs when the weights for all the exemplars labelled in that category approach zero. This represents the listener no longer remembering the category. The listener will cease to produce tokens from that category. A necessary condition for the extinction of a category is that the probability of classifying a sound as that category must approach zero. In this paper we are particularly interested in when there is extinction of all but one category.

This paper is motivated by research in exemplar dynamics carried out by Tupper [15] and Wedel [16]. They both studied the same exemplar dynamic model, but with a subtle difference. In [16], categories were limited to a maximum of 100 stored exemplars, whereas in [15] categories had no limitation on the number of stored exemplars. Tupper demonstrated that if the number of exemplars stored is unlimited, then there is extinction of all but one category. Wedel observed that there is no category extinction in simulations for a certain choice of parameters when the exemplars stored per category is limited to 100. However, Wedel only performed numerical simulations of his model up to 4000 iterations.

https://doi.org/10.1017/jpr.2018.25 Published online by Cambridge University Press

This begs the question, when you limit the number of exemplars to be stored per category, will categories eventually become extinct? In this paper we seek the answer to this question. The models of [15] and [16] are too complicated to investigate rigorously, so we study a simpler model which captures some of their essential features.

In Section 2 we describe our simple exemplar model. Our model depends only on three parameters: the number of categories k, the decay rate λ, and the number of exemplars stored per category N. Two particular cases of this general model will be studied: one where we limit the number of exemplars ($N < \infty$, as in [16]) in Section 3, and another where we do not ($N = \infty$, as in [15]) in Section 4. We prove in both cases that all categories but one will become extinct. In Section 5 we discuss computational results, which demonstrate how limiting the number of exemplars affects the system's evolution. The numerical simulations in this section will help us explain the effect N and λ have on the expected time to extinction.

2. Simple exemplar weight model

In this section we describe a simplified exemplar model. The parameters for the system are the number of categories k, the number of exemplars stored per category N, and the decay rate λ. The listener starts with some exemplars with associated weights in each category, and then receives a stream of new inputs (sounds). The listener in this model will decide how to classify new sounds only using the total weights of the exemplars in each category. The phonetic information stored in exemplars will not be utilized in the categorization process.

At time n, let $w_{j,m}^n$ be the weight of the mth exemplar, where $m \in \mathbb{N}$, for category j. At time n, these k infinite sequences of real numbers comprise the state of the system. Note that throughout this paper, superscript n is an index referring to time n and not the exponent n. Let N be the maximum number of exemplars per category the listener is permitted to store. Let $\lambda > 0$ be the decay rate of the weights, so that at each time-step n, the weights of old memories will decay by a factor of $\beta = \mathrm{e}^{-\lambda}$. New exemplars are given a weight $W_0 = 1$. Additionally, when $N < \infty$, if there are $N + 1$ exemplars in a category with nonzero weight upon adding a new exemplar, then the exemplar with the lowest weight is discarded.

We assume that exemplars are ordered by weight at all times, so that $0 \le w_{j,m+1}^n \le w_{j,m}^n \le 1$ for all $n \in \mathbb{N}_0$, $j \in \{1, \ldots, k\}$, and $m < N$. The initial conditions of the weights are nonrandom, and can be any such that $0 \le w_{j,m}^0 \le 1$ for all $j \in \{1, \ldots, k\}$ and $m \le N$. At least one of the weights in one category must be nonzero, and if $N < \infty$ then $w_{j,m}^0 = 0$ for all $j \in \{1, \ldots, k\}$ and $m > N$.

Let $W_j^n := \sum_{m=1}^N w_{j,m}^n$ be the total weight of exemplars in category $j \in \{1, \ldots, k\}$, and $W_{\mathrm{tot}}^n := \sum_{j=1}^k W_j^n$ be the total weight of all exemplars.

Let x_n be the category we classify as the nth sound at time n. For example, $x_n = j$ means we classified the nth sound as category j. We let the probability of classifying the nth sound at category j ($x_n = j$), given the state of the system in the previous time-step, be $W_j^n / W_{\mathrm{tot}}^n$. This classification procedure is the Luce choice rule [6]. As such, the categorization of sounds depends only on the weights of the exemplars, unlike other models where the phonetic information stored in exemplars is used to classify sounds.

To aid in the analysis of our model, we define a filtration to which the processes $\{w_{j,m}^n\}_{n \ge 0}$ are adapted. First, let $\mathcal{F}$ be the σ-field generated by all random variables in the model. We then define the sequence of σ-fields $\mathcal{F}_n$ for $n \ge 0$ by

$$\mathcal{F}_n = \sigma(w_{j,m}^q,\ 0 < q \le n,\ j \in \{1, \ldots, k\},\ m \in \mathbb{N}). \tag{1}$$

This sequence of σ-fields forms a filtration since, for all n, $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}$; see [2, p. 458].

https://doi.org/10.1017/jpr.2018.25 Published online by Cambridge University Press

Another way to describe the process is as follows. At time-step n,

$$\mathbb{P}(x_n = j \mid \mathcal{F}_n) = \frac{W_j^n}{W_{\text{tot}}^n} \quad \text{for each j} \in \{1, \ldots, k\}.$$

If $x_n = j$ then we have the following.

- Let $w_{j,m+1}^{n+1} = \beta w_{j,m}^n$ for all $m < N$ and $w_{j,1}^{n+1} = W_0 = 1$. If $N < \infty$, the exemplar corresponding to the Nth position of category j from the previous time-step will be discarded: Thus, if $N < \infty$, we let $w_{j,N+1}^{n+1} = 0$.
- For all $i \neq j$ and $m \in \{1, \ldots, N\}$, let $w_{i,m}^{n+1} = \beta w_{i,m}^n$.

We devote the next two sections to proving that the model just described always results in the extinction of all but one category. In Sections 3 and 4 we will look at the cases where $N < \infty$ and $N = \infty$, respectively.

3. Finite stored exemplars model

In this section we will show that when $N < \infty$, all but one category will become extinct with probability 1. That is, it will be proved that with probability 1, there exists an M and a j such that $x_n = j$ for all $n \geq M$.

In the following lemma we prove that if we classify p consecutive sounds as category j, then it only increases the probability of the next sound being classified as category j.

Lemma 1. *If $N < \infty$ and $\mathcal{F}_n$ is the σ-field defined by (1) then*

$$\mathbb{P}(x_{n+p} = j \mid x_{n+p-1} = j, \ldots, x_n = j, \mathcal{F}_n) \geq \mathbb{P}(x_n = j \mid \mathcal{F}_n)$$

almost surely (a.s.) for all $p \in \mathbb{N}_0$ and $j \in \{1, \ldots, k\}$.

Proof. We will prove this lemma using induction. Let $S(p)$ be the statement that

$$\mathbb{P}(x_{n+p} = j \mid x_{n+p-1} = j, \ldots, x_n = j, \mathcal{F}_n) \geq \mathbb{P}(x_n = j \mid \mathcal{F}_n) \quad \text{a.s.}$$

We want to prove that $S(p)$ holds for all $p \in \mathbb{N}_0$.

The initial statement $S(0)$, which is $\mathbb{P}(x_n = j \mid \mathcal{F}_n) \geq \mathbb{P}(x_n = j \mid \mathcal{F}_n)$, holds since the two sides are equal.

Now we assume that the inductive hypothesis $S(p)$ holds; that is,

$$\mathbb{P}(x_{n+p} = j \mid x_{n+p-1} = j, \ldots, x_n = j, \mathcal{F}_n) \geq \mathbb{P}(x_n = j \mid \mathcal{F}_n).$$

We want to show that $S(p+1)$ holds. If $\{x_{n+p} = j, \ldots, x_n = j\}$ then $W_j^{n+p+1} = \beta W_j^{n+p} + 1 - \beta w_{j,N}^{n+p}$ and $W_{\text{tot}}^{n+p+1} = \beta W_{\text{tot}}^{n+p} + 1 - \beta w_{j,N}^{n+p}$. We can show that, via simple algebra and using the facts that $\beta^{-1}(1 - \beta w_{j,N}^{n+p}) > 0$ and $W_j^{n+p} \leq W_{\text{tot}}^{n+p}$,

$$\begin{aligned} \mathbb{P}(x_{n+p+1} = j \mid x_{n+p} = j, \ldots, x_n = j, \mathcal{F}_n) &= \frac{W_j^{n+p} + \beta^{-1}(1 - \beta w_{j,N}^{n+p})}{W_{\text{tot}}^{n+p} + \beta^{-1}(1 - \beta w_{j,N}^{n+p})} \\ &\geq \frac{W_j^{n+p}}{W_{\text{tot}}^{n+p}} \quad \text{a.s.} \end{aligned} \tag{2}$$

The right-hand side of (2) is equal to $\mathbb{P}(x_{n+p} = j \mid x_{n+p-1} = j, \ldots, x_n = j, \mathcal{F}_n)$, which implies, by the induction hypothesis, that statement $S(p+1)$ holds. □

https://doi.org/10.1017/jpr.2018.25 Published online by Cambridge University Press

Define A_n to be the event that we only classify input sounds as a single category from time-step n onwards. More precisely,

$$A_n = \{\text{there exists } j, x_m = j \text{ for all } m \geq n\}. \tag{3}$$

The event A_n will be important throughout this section.

Lemma 2. *If $N < \infty$ and $\mathcal{F}_n$ is the σ-field defined by (1) then there exists a $Q > 0$ such that $\mathbb{P}(A_n \mid \mathcal{F}_n) \geq Q$ for all n, where A_n is the event defined by (3).*

Proof. At time-step n, there must exist a category $c \in \{1, \ldots, k\}$ such that $\mathbb{P}(x_n = c \mid \mathcal{F}_n) \geq k^{-1}$. By Lemma 1, we obtain

$$\begin{aligned}
&\mathbb{P}(x_{n+N-1} = c, \ldots, x_{n+1} = c, x_n = c \mid \mathcal{F}_n) \\
&\quad = \prod_{p=0}^{N-1} \mathbb{P}(x_{n+p} = c \mid x_{n+p-1} = c, \ldots, x_n = c, \mathcal{F}_n) \\
&\quad \geq k^{-N}.
\end{aligned} \tag{4}$$

If $x_q = c$ for $n \leq q \leq n + N - 1$ then $W_c^{n+N} = \sum_{q=0}^{N-1} \beta^q$, and if we continue to categorize $x_q = c$ for $q > n + N - 1$, the weight for category c will stay constant. Let $\Omega = \sum_{q=0}^{N-1} \beta^q$. Upon inspection, it is apparent that $W_i^n \leq \Omega$ for all i and $n > N - 1$, since it is the maximum total weight a category can have after there have been at least N time steps.

If $x_p = c$ for $n \leq p \leq n + q$, where $q \geq N - 1$, then

$$W_{\text{tot}}^{n+q} = \sum_{p=1}^{k} W_p^{n+q} = W_c^{n+q} + \sum_{p \neq c} W_p^{n+q} \leq \Omega + \sum_{p \neq c} \beta^q \Omega < \Omega(1 + k\beta^q).$$

Let $G_{n,N} = \{x_{n+N-1} = c, \ldots, x_{n+1} = c, x_n = c\}$. The probability of categorizing the next sound as c, given that we have only categorized as c since time-step n and have at least done so N times in a row, can be bounded below as

$$\begin{aligned}
\mathbb{P}(x_{n+N-1+q} = c \mid x_{n+N-2+q} = c, \ldots, x_{n+N} = c, G_{n,N}, \mathcal{F}_n) &= \frac{\Omega}{W_{\text{tot}}^{n+N-1+q}} \\
&\geq \frac{\Omega}{\Omega(1 + k\beta^q)} \\
&= 1 - \frac{k\beta^q}{1 + k\beta^q}
\end{aligned} \tag{5}$$

for all $n, q, N > 0$. Note that we used the fact that $W_j^{n+N-1+q} = \Omega$, since the event $G_{n,N}$ had already occurred.

Utilizing (4) and (5),

$$\begin{aligned}
&\mathbb{P}(x_m = c \text{ for all } m \geq n \mid \mathcal{F}_n) \\
&\quad = \mathbb{P}\left(\bigcap_{q=0}^{\infty} \{x_{n+q} = c\} \;\middle|\; \mathcal{F}_n \right) \\
&\quad = \mathbb{P}\left(\bigcap_{q=1}^{\infty} \{x_{n+N-1+q} = c\} \;\middle|\; G_{n,N}, \mathcal{F}_n \right) \mathbb{P}(G_{n,N} \mid \mathcal{F}_n)
\end{aligned}$$

https://doi.org/10.1017/jpr.2018.25 Published online by Cambridge University Press

$$\geq k^{-N} \prod_{q=1}^{\infty} \mathbb{P}(x_{n+N-1+q} = c \mid x_{n+N-2+q} = c, \ldots, x_{n+N} = c, G_{n,N}, \mathcal{F}_n)$$

$$\geq k^{-N} \prod_{q=1}^{\infty} \left(1 - \frac{k\beta^q}{1 + k\beta^q}\right). \tag{6}$$

From Theorem 15.5 of [13], the product in (6) is strictly greater than 0 if and only if

$$\sum_{q=1}^{\infty} \frac{k\beta^q}{1 + k\beta^q} < \infty.$$

By the ratio test we know this series is convergent. Therefore, there is a $Q > 0$ such that $\mathbb{P}(x_m = c \text{ for all } m \geq n \mid \mathcal{F}_n) \geq Q > 0$.

Since

$$\mathbb{P}(\text{there exists: } j, x_m = j \text{ for all } m \geq n \mid \mathcal{F}_n) \geq \mathbb{P}(x_m = c \text{ for all } m \geq n \mid \mathcal{F}_n) \geq Q > 0,$$

we obtain the final result. □

Lemma 2 states that the probability of A_n (that is, (3)) occurring, given any event which only depends on the events up to time-step $n-1$, can be bounded below by a constant $Q > 0$. In other words, the probability of x_m being classified as the same category for all $m \geq n$ always has at least a certain probability of happening no matter what occurs before it.

Note that if A_n holds for any value of n, the rest of the categories $i \neq j$ will become extinct. If we prove that $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = 1$ then we have proved there is a.s. extinction of all but one category when $N < \infty$.

Lemma 3. *Let $\mathcal{G}$ be a σ-field, $G \in \mathcal{G}$ such that $\mathbb{P}(G) > 0$, and X be an event. If there exists a $Q > 0$ such that $\mathbb{P}(X \mid \mathcal{G}) \geq Q$ a.s. then $\mathbb{P}(X \mid G) \geq Q$.*

Proof. By the definition of the probability of an event conditioned on a σ-field (see [12, p. 155]),

$$\mathbb{P}(X \mid G) = \frac{\mathbb{P}(X \cap G)}{\mathbb{P}(G)} = \frac{\mathbb{E}[\mathbb{P}(X \mid \mathcal{G})\, \mathbf{1}_G]}{\mathbb{P}(G)} \geq \frac{\mathbb{E}[Q\, \mathbf{1}_G]}{\mathbb{P}(G)} \geq Q,$$

where $\mathbf{1}_A$ is the indicator function on the event A. □

Theorem 1. *When $N < \infty$, all categories but one will become extinct with probability 1; that is, $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = 1$, where A_n is given by (3).*

Proof. The proof utilizes Murphy's law, a general statement proven in [14]. Murphy's law states the following:

> let $(G_n,\ n \geq 1)$ be any sequence of events satisfying the condition $G_n \subseteq G_{n+1}$ for all $n \geq 1$, and let $G = \bigcup_{n=1}^{\infty} G_n$. If $\mathbb{P}(G \mid G_n^c) \geq \varepsilon > 0$ for all $n \geq 1$ then $\mathbb{P}(G) = 1$.

We know that $A_n \subseteq A_{n+1}$ for all $n \geq 1$. Let $A = \bigcup_{n=1}^{\infty} A_n$. By Murphy's law, if we can show that $\mathbb{P}(A \mid A_n^c) \geq \varepsilon > 0$ for all n then $\mathbb{P}(A) = 1$, proving the theorem.

https://doi.org/10.1017/jpr.2018.25 Published online by Cambridge University Press

Let $Y_n = \min\{m \in \mathbb{N}\colon x_{n+m} \neq x_n\}$, and if it is not defined then let $Y_n = \infty$. The event $\{Y_n = m\}$ is a subset of $A_n^c = \{\text{there exists } j > n \text{ such that } x_j \neq x_n\}$ for all n, and $A_n^c = \bigcup_{m>0}\{Y_n = m\}$. Using the fact that the events $\{Y_n = i\}$ and $\{Y_n = j\}$ are disjoint when $i \neq j$, we obtain

$$\begin{aligned}
\mathbb{P}(A \mid A_n^c) &= \frac{\mathbb{P}(A \cap A_n^c)}{\mathbb{P}(A_n^c)} \\
&= \frac{\mathbb{P}(A \cap (\bigcup_{m>0}\{Y_n = m\}))}{\mathbb{P}(A_n^c)} \\
&= \frac{\mathbb{P}(\bigcup_{m>0}\{A \cap \{Y_n = m\}\})}{\mathbb{P}(A_n^c)} \\
&= \frac{\sum_{m>0}\mathbb{P}(A \cap \{Y_n = m\})}{\mathbb{P}(A_n^c)} \\
&\geq \frac{\sum_{m>0}\mathbb{P}(A_{n+m+1} \cap \{Y_n = m\})}{\mathbb{P}(A_n^c)}
\end{aligned}$$

since $A_{n+m+1} \subseteq A$. Using Lemma 2 with Lemma 3 (noting that $\{Y_n = m\}$ is in $\mathcal{F}_{n+m}$), and the fact that $\bigcup_{m>0}\{Y_n = m\} = A_n^c$, we obtain

$$\begin{aligned}
\frac{\sum_{m>0}\mathbb{P}(A_{n+m+1} \cap \{Y_n = m\})}{\mathbb{P}(A_n^c)} &= \sum_{m>0}\mathbb{P}(A_{n+m+1} \mid \{Y_n = m\})\frac{\mathbb{P}(Y_n = m)}{\mathbb{P}(A_n^c)} \\
&\geq Q\sum_{m>0}\mathbb{P}(Y_n = m \mid A_n^c) \\
&= Q\mathbb{P}\left(\bigcup_{m>0}\{Y_n = m\} \,\middle|\, A_n^c\right) \\
&= Q \\
&> 0.
\end{aligned}$$

□

4. Infinite stored exemplars weight model

This section is devoted to studying the special case of the model where the listener stores an infinite number of exemplars, so $N = \infty$. The proof for showing there is a.s. extinction of all but one category in this special case will be different from the previous section.

Let $Z_j^n = \mathbb{P}(x_n = j \mid \mathcal{F}_n) = W_j^n / W_{\text{tot}}^n$, where $\mathcal{F}_n$ is as defined in (1). Note that the combined weight of all categories which are not j is equal to $W_{\text{tot}}^n - W_j^n$. We will first re-describe the model's evolutionary process in terms of W_j^n and W_{tot}^n in order to simplify the proof. The evolutionary process evolves as follows.

- If $x_n = j$ then the total weight of category j becomes $W_j^{n+1} = 1 + W_j^n\beta$, and the total weight of all other categories besides j becomes

$$W_{\text{tot}}^{n+1} - W_j^{n+1} = (W_{\text{tot}}^n - W_j^n)\beta.$$

- If $x_n \neq j$ then the total weight of all categories besides j is

$$W_{\text{tot}}^{n+1} - W_j^{n+1} = 1 + (W_{\text{tot}}^n - W_j^n)\beta,$$

and the total weight of category j is $W_j^{n+1} = W_j^n\beta$.

https://doi.org/10.1017/jpr.2018.25 Published online by Cambridge University Press

We want to prove that there exists a category j such that $Z_j^n \to 1$ a.s. as $n \to \infty$, and for the rest of the categories $q \neq j$, $Z_q^n \to 0$ a.s. We will then show that if $Z_j^n \to 0$ a.s. then $W_j^n \to 0$ a.s. As such we would prove that all categories but one become extinct. In order to prove this result, we require some additional lemmas.

Lemma 4. *Let $Z_j^n = W_j^n / W_{\text{tot}}^n$. If the number of exemplars per category stored is $N = \infty$ then the random variable Z_j^n is a martingale with respect to the filtration $\{\mathcal{F}_n\}_{n\geq 1}$.*

Proof. We know that Z_j^n is $\mathcal{F}_n$-measurable (see [2, p. 68]) and $\mathbb{E}[|Z_j^n|] \leq 1$. Due to the fact that Z_j^{n+1} conditioned on $\mathcal{F}_n$ only depends on the values of W_j^n and W_{tot}^n, we obtain

$$\begin{aligned}
\mathbb{E}[Z_j^{n+1} \mid \mathcal{F}_n] &= \mathbb{E}[Z_j^{n+1} \mid W_j^n, W_{\text{tot}}^n] \\
&= \frac{W_j^n}{W_{\text{tot}}^n}\left(\frac{W_j^n\beta + 1}{W_{\text{tot}}^n\beta + 1}\right) + \frac{W_{\text{tot}}^n - W_j^n}{W_{\text{tot}}^n}\left(\frac{W_j^n\beta}{W_{\text{tot}}^n\beta + 1}\right) \\
&= \frac{W_j^n}{W_{\text{tot}}^n} \\
&= Z_j^n,
\end{aligned}$$

implying that Z_j^n is a martingale with respect to the filtration $\{\mathcal{F}_n\}_{n\geq 1}$; see [2, p. 458]. □

Lemma 5. *There exists a $\gamma \in \mathbb{R}$ depending only on λ and the initial total weight W_{tot}^0 such that $W_{\text{tot}}^n \leq \gamma$ for $i = 1, 2, \ldots, k$ and for all $n \geq 0$.*

Proof. We know that $W_{\text{tot}}^n = \sum_{i=1}^k = W_{\text{tot}}^{n-1}\mathrm{e}^{-\lambda} + 1$ for all realizations. Since $\mathrm{e}^{-\lambda} < 1$, we know that W_{tot}^n converges to $W := (1 - \mathrm{e}^{-\lambda})^{-1}$. Since W_{tot}^n converges monotonically to W,

$$W_{\text{tot}}^n \leq \max\left\{W_{\text{tot}}^0, \frac{1}{1 - \mathrm{e}^{-\lambda}}\right\} = \gamma \quad \text{for all } n.$$

This, in turn, implies the result. □

Using Lemmas 4 and 5, and the martingale convergence theorem (see [2, p. 468]), we are able to prove Theorem 2.

Theorem 2. *If $N = \infty$ then, for all $j \in \{1, \ldots, k\}$, Z_j^n converges a.s. to a random variable Z_j^* a.s. Furthermore, the only values that Z_j^* can be with positive probability are* 0 *and* 1.

Proof. To prove this theorem, we will require an expression for $\text{var}(Z_j^{n+1} \mid \mathcal{F}_n)$, where $\mathcal{F}_n = \sigma(w_{j,m}^q,\ q \leq n,\ j \in \{1, \ldots, k\}, m \in \mathbb{N})$, as in Lemma 4. First, we determine $\mathbb{E}[(Z_j^{n+1})^2 \mid \mathcal{F}_n]$; that is,

$$\begin{aligned}
\mathbb{E}[(Z_j^{n+1})^2 \mid \mathcal{F}_n] &= \frac{W_j^n}{W_{\text{tot}}^n}\left(\frac{W_j^n\beta + 1}{W_{\text{tot}}^n\beta + 1}\right)^2 + \frac{W_{\text{tot}}^n - W_j^n}{W_{\text{tot}}^n}\left(\frac{W_j^n\beta}{W_{\text{tot}}^n\beta + 1}\right)^2 \\
&= \frac{(W_j^n)^2 W_{\text{tot}}^n\beta^2 + 2(W_j^n)^2\beta + W_j^n}{W_{\text{tot}}^n(W_{\text{tot}}^n\beta + 1)^2}.
\end{aligned}$$

https://doi.org/10.1017/jpr.2018.25 Published online by Cambridge University Press

This allows us to calculate the conditional variance,

$$\begin{aligned}\operatorname{var}(Z_j^{n+1} \mid \mathcal{F}_n) &:= \mathbb{E}[(Z_j^{n+1})^2 \mid \mathcal{F}_n] - \mathbb{E}[Z_j^{n+1} \mid \mathcal{F}_n]^2 \\ &= \frac{(W_j^n)^2 W_{\text{tot}}^n \beta^2 + 2(W_j^n)^2 \beta + W_j^n}{W_{\text{tot}}^n (W_{\text{tot}}^n \beta + 1)^2} - \left(\frac{W_j^n}{W_{\text{tot}}^n}\right)^2 \\ &= W_j^n (W_{\text{tot}}^n - W_j^n)(W_{\text{tot}}^n)^{-2} (W_{\text{tot}}^n \beta + 1)^{-2} \\ &= Z_j^n (1 - Z_j^n)(W_{\text{tot}}^n \beta + 1)^{-2}. \end{aligned} \tag{7}$$

By the martingale convergence theorem, since Z_j^n is a submartingale and $\sup_n \mathbb{E}|Z_j^n| \leq 1$, we know that there is a random variable Z_j^* such that $Z_j^n \to Z_j^*$ a.s. This implies that $Z_j^{n+1} - Z_j^n \to 0$ a.s., and we know that $|Z_j^{n+1} - Z_j^n| \leq 2$ for all n. By the dominated convergence theorem (see [12]), this implies that $\mathbb{E}[|Z_j^{n+1} - Z_j^n|^2] \to 0$ as $n \to \infty$.

By Lemma 5, $W_{tot}^n \leq \gamma$ for all n. Since Z_j^n is $\mathcal{F}_n$-measurable and $\mathbb{E}[Z_j^{n+1} \mid \mathcal{F}_n] = Z_j^n$,

$$\begin{aligned}\mathbb{E}[|Z_j^{n+1} - Z_j^n|^2] &= \mathbb{E}[(Z_j^{n+1})^2 - 2Z_j^{n+1} Z_j^n + (Z_j^n)^2] \\ &= \mathbb{E}[\mathbb{E}[(Z_j^{n+1})^2 - 2Z_j^{n+1} Z_j^n + (Z_j^n)^2 \mid \mathcal{F}_n]] \\ &= \mathbb{E}[\operatorname{var}(Z_j^{n+1} \mid \mathcal{F}_n)]. \end{aligned} \tag{8}$$

Using (7) and (8), as well as Lemma 5, we obtain

$$\mathbb{E}[|Z_j^{n+1} - Z_j^n|^2] = \mathbb{E}[Z_j^n (1 - Z_j^n)(W_{\text{tot}}^n \beta + 1)^{-2}] \geq (\gamma\beta + 1)^{-2} \mathbb{E}[Z_j^n (1 - Z_j^n)].$$

Taking the limit as $n \to \infty$ on both sides, we obtain $\mathbb{E}[Z_j^n (1 - Z_j^n)] \to 0$ as $n \to \infty$. Since convergence in L_1 implies convergence in probability (see [11, p. 85]), we have

$$\mathbb{P}(Z_j^n (1 - Z_j^n) < \varepsilon) \to 1 \quad \text{for all } \varepsilon > 0.$$

This implies that there exists a subsequence such that $Z_j^{n_i}(1 - Z_j^{n_i}) \to 0$ a.s.; see [1, p. 7]. As such Z_j^* can only equal 0 or 1, since we know there must exist a Z_j^* such that $Z_j^n \to Z_j^*$ a.s. The proof is complete. □

This brings us to our final result.

Theorem 3. *When $N = \infty$, in the model described in Section 2, all categories but one will become extinct with probability* 1.

Proof. From Lemma 5, we know that $Z_j^n = W_j^n / W_{\text{tot}}^n \geq W_j^n \gamma^{-1} \geq 0$, implying that if $Z_j^n \to 0$ then $W_j^n \to 0$ as well. By Theorem 2, for every category $j \in \{1 \ldots k\}$, $Z_j^n \to Z_j^*$ a.s., where Z_j^* can only be 0 or 1, and we know that $\sum_j Z_j^* = 1$. As such $Z_j^n \to 0$ a.s. for every category j but one. This implies all but one category will become extinct with probability 1. The proof is complete. □

5. Simulations and time to extinction

In the last two sections we proved that extinction of all but one category occurs for our model regardless of the value of N. In this section we will discuss some of the results obtained by computer simulations of the simplified weight model. These simulations will demonstrate

https://doi.org/10.1017/jpr.2018.25 Published online by Cambridge University Press

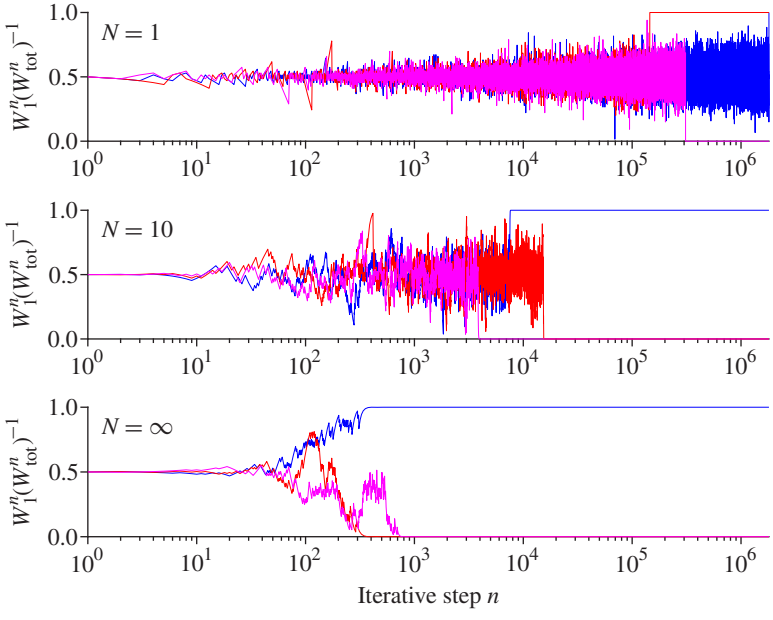

FIGURE 2: Plots of $Z_1^n = W_1^n / W_{\text{tot}}^n$ for single simulations when $k = 2$, $\lambda = 0.06$, and the weight threshold is $10^{-4} W_0$. For each value of N we have plotted Z_1^n against time-step n for three simulations.

how changing the variables N and λ affects how long it takes until there is only one nonextinct category left in the system.

Before discussing the results of our computer simulations, we will explain weight thresholds. Analytically, a category j becomes extinct when $W_j^n \to 0$ as $n \to \infty$. Extinction of all but one category means there exists a category j such that $W_i^n \to 0$ as $n \to \infty$ for all $i \neq j$. When running computer simulations, we cannot possibly know for certain if a category's weight approaches 0, but we do something else to detect if it is most likely to do so. In simulations, once a category's weight goes below a value we call a weight threshold, we assume that the category becomes extinct. The time it takes for all but one of the category's weights to go below the weight threshold will be referred to as the extinction time. In Figures 2 and 3, the number of categories is $k = 2$, so the extinction time is how soon one of the two categories goes extinct.

In Figure 2 we present three simulations each for three separate values of N, where the number of categories is $k = 2$. We show the evolution of the random variable Z_1^n (defined in Section 4) for the values $N = 1$, 10, and ∞. When Z_1^n hits either 0 or 1, the simulation ends, representing that either category 1 or 2 has become extinct, respectively. Upon inspection, we see that the larger N is, the faster categories become extinct.

In Figure 3 we see how the expected extinction time changes based on the values of our decay rate λ, and the limitation on the number of exemplars N, when the number of categories is $k = 2$. The expected value for the extinction time is found by averaging over 1000 simulations for each value of N and λ. As N decreases, we observe, as in Figure 2, that the extinction time increases. Likewise, as λ decreases the extinction time increases as well.

It is straightforward to explain how λ affects the extinction time, but the explanation for the effect of N is more subtle. To help understand the effect of N on the extinction time, we will consider two examples. For both examples, let $\beta = \frac{1}{2}$, $k = 2$ (two categories), and the initial

https://doi.org/10.1017/jpr.2018.25 Published online by Cambridge University Press

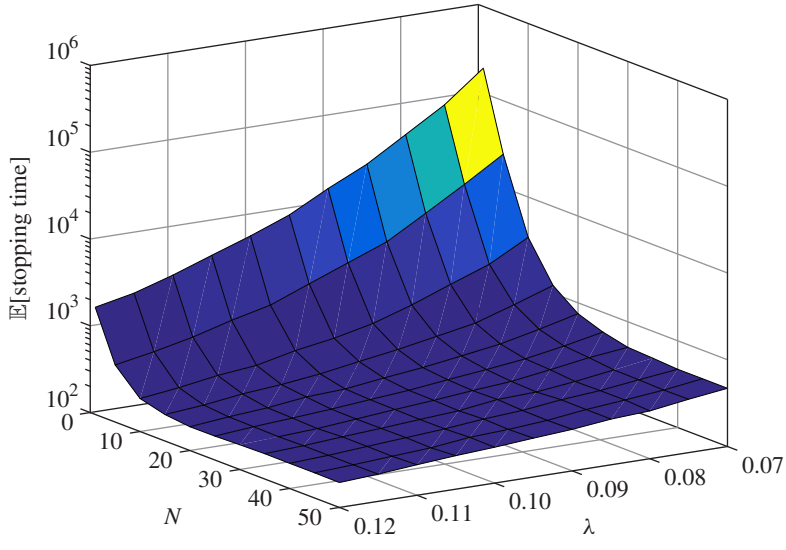

FIGURE 3: A plot of the expected extinction time as we change variables N and λ. We use a weight threshold equal to $10^{-4}W_0$.

weight of the first two exemplars in each list is $W_0 = 1$, while the rest of the exemplar weights are 0.

(i) First consider the case where $N = 2$. If $x_0 = 2$, the weights of category 2 will be $w^2_{2,1} = 1$ and $w^2_{2,2} = \frac{1}{2}$, and the weights of category 1 will be $w^2_{1,1} = w^2_{1,2} = \beta = \frac{1}{2}$. This implies the probability that $x_1 = 2$, given that $x_0 = 2$, is 60%.

(ii) Now consider the case where $N = \infty$. If $x_0 = 2$ then the total weight of categories 1 and 2 respectively will be $W^n_1 = 2\beta = 1$ and $W^n_2 = 2\beta + 1 = 2$. This implies the probability that $x_1 = 2$, given that $x_0 = 2$, is approximately 66.7%.

It is more probable, when $N = \infty$, for a category to be consecutively categorized. When $N = 2$, it is rarer for the exemplar weights to decay close to 0 than when $N = \infty$. This demonstrates why limiting the number of exemplars makes extinction take longer. When $N = \infty$, exemplars getting stored in a category consecutively adds comparatively more weight to the category. This explains the effect of N on the extinction time, as seen in Figures 2 and 3.

The behaviour of the expected extinction time increasing as λ decreases is much easier to explain. The weights are decaying slower, so it will take longer for the weights to approach 0. If $\lambda = 0$ then there would be no decay and thus no category extinction. Because of this, as $\lambda \to 0$, the expected extinction time will asymptotically approach infinity.

6. Discussion

The model studied in this paper is simpler than the ones studied by Tupper [15] and Wedel [16], but it helps explain the behaviour we see in these models. Changing N in our model does not affect whether all categories but one eventually become extinct, but it does affect the time it takes to do so. Our results agree with the extinction result demonstrated in [15] for $N = \infty$. However, our work suggests that the model studied in [16] will eventually show the

https://doi.org/10.1017/jpr.2018.25 Published online by Cambridge University Press

same behaviour but on a longer time scale. This longer time scale may explain why category extinction was not observed in Wedel's simulation [16].

One natural direction we can take in future research is to apply our model to real-world data. For example, in Figure 1 we illustrated the evolution of the usage of two spellings of the word cider over 200 years [7]. The archaic spelling 'cyder' becomes extinct close to the year 1980. Using the corpus of digitized texts put together in [7], one could determine what values of N and λ best model this type of data.

Acknowledgements

This work was supported by an NSERC Discovery Grant, an NSERC Discovery Accelerator Supplement, and a Tier 2 Canada Research Chair.

References

[1] Bhattacharya, R. and Waymire, E. C. (2007). *A Basic Course in Probability Theory*. Springer, New York.
[2] Billingsley, P. (1995). *Probability and Measure*, 3rd edn. John Wiley, New York.
[3] Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Stud. Second Language Acquisition* **24,** 215–221.
[4] Jäger, G. (2008). Applications of game theory in linguistics. *Language Linguistics Compass* **2,** 406–421.
[5] Johnson, K. (1997). Speech perception without speaker normalization: an exemplar model. In *Talker Variability in Speech Processing*, Academic Press, San Diego, CA, pp. 145–165.
[6] Luce, R. D. (2005). *Individual Choice Behavior: A Theoretical Analysis*. Dover, Mineola, NY.
[7] Michel, J. B. *et al.* (2011). Quantitative analysis of culture using millions of digitized books. *Science* **331,** 176–182.
[8] Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *J. Experimental Psychology General* **115,** 39–57.
[9] Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *J. Experimental Psychology Learning Mem. Cognition* **14,** 54–65.
[10] Pierrehumbert, J. B. (2001). Exemplar dynamics: word frequency, lenition, and contrast. In *Frequency and the Emergence of Linguistic Structure*, John Benjamins, Amsterdam, pp. 137–157.
[11] Romano, J. P. and Siegel, A. F. (1986). *Counterexamples in Probability and Statistics*. Wadsworth and Brooks/Cole, Monterey, CA.
[12] Rosenthal, J. S. (2000). *A First Look at Rigorous Probability Theory*. World Scientific, River Edge, NJ.
[13] Rudin, W. (1987). *Real and Complex Analysis*, 3rd edn. McGraw-Hill, New York.
[14] Steel, M. (2015). Reflections on the extinction–explosion dichotomy. *Theoret. Pop. Biol.* **101,** 61–66.
[15] Tupper, P. F. (2015). Exemplar dynamics and sound merger in language. *SIAM J. Appl. Math.* **75,** 1469–1492.
[16] Wedel, A. (2012). Lexical contrast maintenance and the organization of sublexical contrast systems. *Language Cognition* **4,** 319–355.
[17] Winter, B. and Wedel, A. (2016). The co-evolution of speech and the lexicon: the interaction of functional pressures, redundancy, and category variation. *Topics Cognitive Sci.* **8,** 503–513.