Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model

Supplementary Materials

Simon Todd, Janet B. Pierrehumbert, Jennifer Hay

Contents

S1 Model details

In this section, we present the equations and technical details underlying the computational model presented in Section 3 of the paper, and make fine-grained comparisons with previous exemplar dynamics models with respect to these details.

S1.1 Initial data

The model is initialized by providing an initial exemplar cloud for each category, i.e. a distribution of exemplars across a set of types. There are three components to the initial exemplar cloud for each category:

- A set of types, whose frequencies follow a specified type-frequency distribution.
- A set of exemplars, whose acoustic values follow a specified acoustic distribution. The total number of exemplars is equal to the sum of type frequencies for the category.
- An assignment of exemplars to types, such that a given type of frequency f has f exemplars, in accordance with the multiple-trace hypothesis (Hintzman & Block, 1971).

We generated a distinct set of 92 types for each category, so that our system did not include minimal pairs. We used the same type-frequency distribution for each category. This distribution was based on the distribution of word log-frequencies in COCA: The Corpus of Contemporary American English (Davies, 2008-) (see Figure S1).

We chose the log-transformation of corpus frequency because it reflects the fact that participants underestimate the frequency of common words (Begg, 1974), and is consistent with the "negatively accelerated, increasing relation between represented and actual frequency" observed by Nosofsky (1991, p. 15). While many other transformations are also consistent with this observation, the log-transformation is widely used in processing models and empirical studies assessing a relationship between word frequency and behavior (in terms of both behavioral response properties – e.g. reaction time and categorization probability – and word realization properties – e.g. duration and acoustic quality), both for words in isolation (e.g. Murray & Forster, 2004, and studies cited therein) and for words in context (e.g. Smith & Levy, 2013, and studies cited therein). Importantly, log-frequencies are used in all of the studies of word frequency effects in sound change that our model attempts to explain.

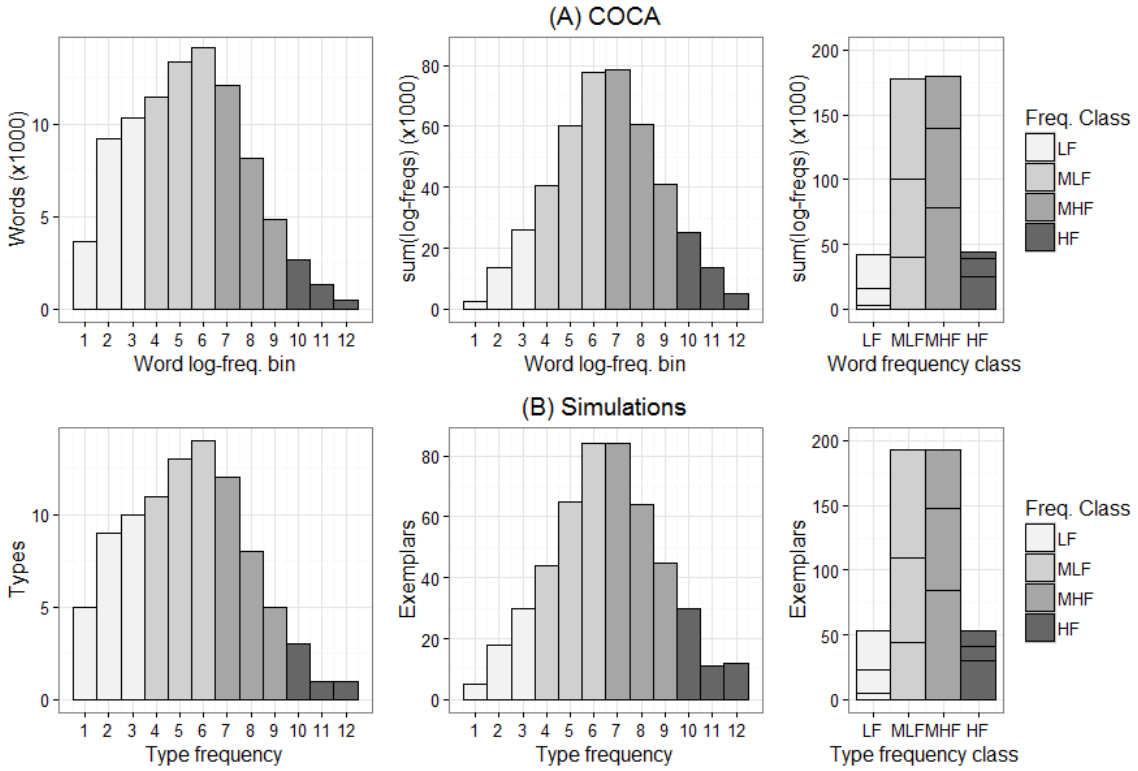

Figure S1: (A) The distribution of word frequencies in COCA (Davies, 2008-). Panels show the number of unique words (left) and the sum of word log-frequencies (middle) in each word log-frequency bin, and the sum of word log-frequencies in each of four word frequency classes (right), from "low-frequency" (bins 1–3; lightest gray) to "high-frequency" (bins 10–12; darkest gray). We calculated log-frequencies using the natural logarithm and binned them by rounding up, then excluded words occurring extremely infrequently (log-frequency 0) or extremely frequently (log-frequency > 12). (B) We modeled the distribution of type frequencies in our simulations on the distribution observed in COCA, equating words with types, and type frequency (and thus number of exemplars per type, following the multiple-trace hypothesis (Hintzman & Block, 1971)) with word log-frequency. This yielded notable symmetry across frequency classes (right panel): there are as many exemplars of low-frequency types as there are of high-frequency types.

As a consequence of using corpus log-frequencies, we obtained the same number of exemplars of high-frequency types (with few types and many exemplars per type) as exemplars of low-frequency types (with many types and few exemplars per type). This means that, in our system, the set of exemplars of all high-frequency types and the set of exemplars of all low-frequency types are both updated at the same rate[1]; any observed effects of type frequency are thus based on differences in the way that types are processed, not on differences in sheer quantity of exemplars.

For each category, we generated a set of 492 exemplars with acoustic values following a raised-cosine distribution. We used a raised-cosine distribution because its short tails are less susceptible to iterated sampling error than the long tails of a normal distribution, and thus are more robust to the effects of discriminability and typicality evaluation. We sampled from this distribution in such a way that the sample of exemplars of high-frequency types was identical to the sample of low-frequency types (and similarly for mid-high- and mid-low-frequency types). This sampling strategy means that there is no frequency asymmetry in the initial conditions of our model; any

[1]In a given period of time, a given high-frequency type is produced and perceived more than a given low-frequency type, but high-frequency types *in aggregate* are produced and perceived the same number of times as low-frequency types *in aggregate*, since there are many fewer high-frequency types than low-frequency types (Zipf, 1935).

Table S1: Initial exemplar distribution statistics for type frequency classes.

Freq.	N	Mean	SD	Skew	Ex. Kurtosis
(All)	492	−0.013	0.998	0.109	−0.717
HF	53	−0.009	0.999	0.097	−0.770
MHF	193	−0.013	0.998	0.112	−0.702
MLF	193	−0.013	0.998	0.112	−0.702
LF	53	−0.009	0.999	0.097	−0.770

observed effects of type frequency over time are thus based on differences in the way that types are processed, not on differences in initial conditions. Finally, we assigned identical (but displaced) initial exemplar sets to each category, to ensure that any observed differences between categories are due to category interaction and not initial conditions.

The process we used to sample the initial exemplar sets was as follows. First, we drew 246 values from the raised cosine distribution whose probability density function is given in Equation (S1) and rounded them to the nearest 0.1.

$$r(v) = \begin{cases} \frac{c}{2}\left(1 + \cos\left(c\pi v\right)\right) & \text{if } v \in \left[\frac{-1}{c}, \frac{1}{c}\right] \\ 0 & \text{otherwise} \end{cases} \tag{S1}$$

where

$$c = \sqrt{\frac{1}{3} - \frac{2}{\pi^2}} \tag{S2}$$

We re-drew values until we obtained a sample with mean approximately equal to zero, standard deviation approximately equal to 1, and low skewness. We then split this sample into two subsets of 53 and 193, ensuring that the statistics for each subset were approximately equal to the statistics across the entire sample. We made two copies of each subset and designated them to classes of types on the basis of type frequency; we designated a copy of the first subset (with 53 exemplars) for each of the high- and low-frequency classes, and a copy of the second subset (with 193 exemplars) for each of the mid-high- and mid-low-frequency classes. The statistics for the sample and frequency-class subsets are given in Table S1.

We used this distribution of acoustic values as the basis of all simulations with the model, adjusting it as required by the parameters of the simulation. To create initial categories of width σ, we scaled the acoustic value of every exemplar by σ. To create an initial category distance of μ between the categories, we subtracted μ from the (scaled) acoustic value of each of the exemplars in the Pusher. Following any adjustments, we re-rounded the acoustic values to the nearest 0.1.

In each run of the model, we assigned the exemplars for a given frequency class to types in that frequency class at random, ensuring that a given type of frequency f had f exemplars. This random assignment means that the results of many runs of the model with a given set of parameter values reflect the dynamics expected on average under that set of parameter values, independent of the effects of initial allocation of exemplars to types.

S1.2 Model processes

Each iteration of the model consists of a single token being produced and submitted to perceptual evaluation (and thus potentially stored). In this section, we describe the implementation details for the processes making up an iteration, and we compare the details in our model to those in other exemplar-based models.

S1.2.1 Type Selection

In type selection, a type (from either category) is chosen at random based on its frequency. The probability of choosing type T_k, of frequency f_k, is given by Equation (S3).[2]

$$P(T_k) = \frac{f_k}{\sum_j f_j} \tag{S3}$$

S1.2.2 Target Selection

In target selection, an exemplar of the selected type is chosen at random, uniformly, and used to provide an acoustic target for the production. The probability of choosing exemplar j of type T_k, with acoustic value $x_{j,k}$, is given by Equation (S4).

$$P(v = x_{j,k}|T_k) = \frac{1}{f_k} \tag{S4}$$

S1.2.3 Bias

If the target type is a member of the Pusher category, then the addition of bias adds β (a parameter) to the target v, yielding a new target v'. If the target type is a member of the Pushee category, no bias is added. This is exemplified in Equation (S5).

$$v' = \begin{cases} v + \beta & \text{if target is Pusher} \\ v & \text{if target is Pushee} \end{cases} \tag{S5}$$

The function of bias is to enforce sustained category interaction and promote long-term movement in one direction. Thus, bias itself does not *cause* categories to interact, but rather gives categories sustained opportunities to interact. In Section S3.3.2, we show that simulations without bias exhibit decreasing category interaction over time.

Our treatment of bias as systematic, i.e. applied to all tokens (of the Pusher) equally, follows that presented by Pierrehumbert (2001, 2002). The major downside to this treatment is that the bias is unconstrained and continues acting in the same way throughout the simulation, generating perpetual category movement. Other authors (Wedel, 2006; Wedel & Fatkullin, 2017; Sóskuthy, 2013; Tupper, 2015) use instead a bias that applies to tokens differentially, based on their distance from some fixed attractor point. This alternative treatment places constraints on the movement induced by the bias, causing movement to cease when the Pusher reaches the attractor. While it is easy to understand how such an attractor may arise in the case of leniting biases (i.e. through the minimization of articulatory effort), it is harder to understand how an attractor may arise in sound change more generally, assuming that it is not something the speaker can agentively establish. Our chosen treatment of systematic bias may be seen as a convenient way to sidestep

[2]The type frequencies underlying production are the same as those underlying perception, and thus reflect real-world log-frequency. See Appendix A.3 of the paper for discussion.

this issue. We point out that the generation of perpetual movement under our treatment is a reflex of the simplicity of the modeling environment: with the inclusion of additional repellers in the system (provided by other categories and/or articulatory limits), movement would no longer be unconstrained (Sóskuthy, 2013). We further point out that our treatment is almost equivalent to an attractor-based treatment in which the attractor is sufficiently far from the Pusher's boundary with the Pushee.

S1.2.4 Imprecision

Under imprecision, random noise n is added to the target v', yielding a final target v'' for the transmitted token, as shown in Equation (S6). n is a single sample from a normal distribution with standard deviation ι (a parameter); the larger ι, the more the target may deviate.

$$v'' = v' + n \qquad n \sim \mathcal{N}(0, \iota^2) \tag{S6}$$

The final target v'' is rounded to the nearest 0.1 before the token is transmitted.

The function of imprecision is to allow a discrete set of exemplars to generate a continuous distribution over the perceptual-acoustic space from which targets can be sampled in production. In this way, imprecision allows for novelty in production targets.

The use of token-wise imprecision generates a non-parametric sampling distribution. This approach is standard in exemplar dynamics models, but other approaches are also possible. For example, Harrington et al. (2018) generate a parametric sampling distribution by inferring a Gaussian distribution over all exemplars of a category. A parametric approach forces all exemplar distributions to have a common shape, with fixed kurtosis and zero skewness. This enforcement makes meeting a model desideratum of shape maintenance almost trivial, which is advantageous; however, it doesn't allow for the modeling of distributions that differ substantially from the parametric (Gaussian) shape.

S1.2.5 Activation

Upon transmission of the token, all exemplars (of both categories) are activated to some extent, according to their distance from the token in the perceptual-acoustic space. The activation A_i of category C_i is given by the sum of the activations of exemplars belonging to that category, as shown in Equation (S7).

$$A_i = \sum_{x \in C_i} w_a(v'' - x) \tag{S7}$$

The degree to which the token activates each exemplar is provided by a Gaussian window w_a with width α (a parameter), as shown in Equation (S8). Exemplars that are very near the token are given activations close to 1, while exemplars that are very far away are given activations close to 0. Increasing α causes exemplars within a wider radius to be given non-negligible activations.

$$w_a(d) = \exp\left(\frac{-d^2}{2\alpha^2}\right) \tag{S8}$$

Most previous exemplar-based models of regular sound change have used a rectangular (Pierrehumbert, 2001, 2002; Ettlinger, 2007) or exponential (Wedel, 2006, 2012; Wedel & Fatkullin, 2017) activation window (though note the use of a Gaussian window by Sóskuthy (2013), by appeal to common practice in kernel density estimation, a statistical technique with the same mathematical

underpinnings as exemplar-based modeling (Ashby & Alfonso-Reese, 1995)). Our use of a Gaussian window here is motivated by discussion in the psychological literature of an equivalent parameter (p) in exemplar-based models of categorization using Multi-Dimensional Scaling. For example, Nosofsky (1985) found that asymptotic human categorization data (i.e. highly successful categorization which accesses pre-learnt structures) is better modeled with a Gaussian activation window than an exponential one, and Shepard (1958) developed an underlying process model predicting that a Gaussian activation window should arise under cases of infrequent feedback of categorization correctness, while an exponential window should arise under continuous feedback (which arguably does not occur in language, at least directly). The Gaussian window also has the practical advantage that it is *smooth*, whereas the rectangular and exponential windows are not (they contain jumps and a sharp peak, respectively); this ensures that the activation fields obtained in the modeling process are also smooth, even when exemplar distributions are sparse.

S1.2.6 Identification

Since we assume no minimal pairs and perfect transmission of the phonological frame, the intended category C_i is always able to be accurately identified. Whether or not the exemplar is stored is determined by the extent to which it is discriminable as a member of the intended category as opposed to the other category, and by the extent to which it is typical of the intended category. Both of these are assessed by probabilistic evaluations.

S1.2.7 Discriminability Evaluation

The probability of passing the discriminability evaluation is determined by comparing the ratio of category activations (intended category activation, A_i, divided by other category activation, A_o) to the discriminability threshold, δ (a parameter). When the ratio is equal to the threshold, the probability of passing the evaluation is 0.5, as shown in Equation (S9). As the ratio grows relative to the threshold, so too does the probability of passing the evaluation. This means that decreasing δ increases the probability of passing the discriminability threshold, and thus increases discriminability.

$$P(\text{pass discriminability evaluation}|A_i, A_o) = \frac{\frac{A_i}{A_o}}{\frac{A_i}{A_o} + \delta} \tag{S9}$$

The formulation of discriminability evaluation in Equation (S9) is equivalent to an application of the Generalized Context Model (Nosofsky, 1986), which extends the application of Luce's Choice Rule (Luce, 1959) over category activations in the Context Model (Medin & Schaffer, 1978) by incorporating category response biases. Here, the bias towards the intended category C_i is $1/\delta$ and the bias towards the other category C_o is 1, as shown by Equation (S10).

$$P(\text{pass discriminability evaluation}|A_i, A_o) = \frac{\frac{1}{\delta} \cdot A_i}{\frac{1}{\delta} \cdot A_i + 1 \cdot A_o} \tag{S10}$$

This equivalence allows for an alternative interpretation of discriminability evaluation, as the act of categorizing the input. Under this interpretation, categorizations that are not sufficiently supported by context – in this case, yielding types corresponding to non-words – are blocked from storage.

Note from Equation (S10) that δ has a multiplicative effect on activations, not an additive effect as in the Logogen model (Morton, 1969). This means that the activation of an exemplar of a high-frequency type is not on average higher than the activation of an exemplar of a low-frequency type, independent of acoustic value; rather, a token of a high-frequency type garners more activation than an otherwise identical token of a low-frequency type (i.e. one with the same acoustic value).

The notion of being unlikely to store tokens that have limited discriminability is also seen in models presented by Wedel (2006, 2012). There, tokens are categorized probabilistically according to the Context Model (Equation (S10) with $\delta = 1$), and the result is stored with probability equal to the categorization probability. This means that tokens that activate both categories to comparable extents, i.e. tokens with limited discriminability, are unlikely to be robustly recognized and stored, just as in the model presented in this paper.

Some models make a stronger assumption that tokens with low discriminability may never be stored. For example, in the model presented by Harrington et al. (2018), tokens are assigned to the category with maximum likelihood and only stored if that category contains a consistent type (i.e. a word with the same phonological frame as the token). In the absence of minimal pairs, a low-discriminability token – with higher acoustic similarity to a nonword type than to the intended type – will never be stored. Because maximum likelihood categorization creates a hard boundary at the intersection point of two categorization probability distributions, it prevents category overlap (Tupper, 2015; Wedel & Fatkullin, 2017).

S1.2.8 Typicality Evaluation

The probability of passing the typicality evaluation is determined by comparing the activation of the intended category, A_i (normalized for the number of exemplars of the category, N_i), to the typicality threshold, τ (a parameter). When the activation is equal to the threshold, the probability of passing the evaluation is 0.5, as shown in Equation (S11). As the activation grows relative to the threshold, so too does the probability of passing the evaluation. This means that decreasing τ increases the probability of passing the typicality threshold, and thus decreases sensitivity to typicality.

$$P(\text{pass typicality evaluation}|A_i) = 1 - \exp\left(-\ln 2 \cdot \frac{A_i}{N_i \tau}\right) \tag{S11}$$

The formulation of typicality evaluation in Equation (S11) is inspired the Complete Set Model of Busemeyer et al. (1984). In this model, a "junk" category competes with established categories in the classification of a token; when the token does not yield sufficient activation, it is discarded as junk. The junk category has no exemplar basis of representation and thus is not included in the application of Luce's Choice Rule (i.e. in the equivalent of Equation (S10)); instead, it discounts the probability mass of each category (as derived from Luce's Choice Rule) by a scale factor. As shown in Equation (S12), this is equivalent to a two-stage process where the scale factor represents the probability associated with a junking decision that is contingent on categorization.

$$P(\text{member of } C_i \text{ and not junk}) = P(\text{member of } C_i) \cdot P(\text{not junk}|\text{member of } C_i) \tag{S12}$$

We equate this post-categorization junking decision with typicality evaluation. Busemeyer et al. (1984) assume that junking is independent of (and thus potentially precedes) categorization, with the probability of junking decreasing exponentially with total activation across all categories. We keep the same form for our treatment of typicality evaluation (Equation (S11)), but assume that

only the activation of the chosen (intended) category contributes. This follows from our treatment of typicality evaluation as occurring after identification and discriminability evaluation and hence assessing the extent to which the token is typical for the category to which it has been confidently assigned.

In our model, typicality evaluation generates a force that squeezes each category toward its mode. This is critically different from previous models (e.g. Pierrehumbert, 2001, 2002; Wedel, 2004, 2006, 2012; Tupper, 2015; Wedel & Fatkullin, 2017), in which the equivalent force (due to *entrenchment*; see Section S4) squeezes each category toward its mean. We believe that squeezing toward the mode is superior to squeezing toward the mean, for two reasons. Consider the case of partially overlapping categories with short tails in the overlapping region (as created by the discriminability evaluation). Firstly, since the mode of each category is located closer to the overlapping region than the mean, squeezing toward the mode will push categories away from each other less than squeezing toward the mean. Thus, squeezing toward the mode will maintain category overlap better than squeezing toward the mean, in line with the model desideratum. Secondly, if the squeezing force grows superlinearly with the distance from the center (mode or mean), then squeezing toward the mode will shorten the short tail less than squeezing toward the mean, since the short tail is closer to the mode than it is to the mean (and vice-versa for the long tail). Thus, squeezing toward the mode will resist increasing category skewness better than squeezing toward the mean, in line with the model desideratum for maintenance of shape.[3]

S1.2.9 Storage

If the token passes both the discriminability evaluation and the typicality evaluation, it is stored and overwrites a random exemplar of the same type in the exemplar space. All exemplars are stored with the same strength, which does not decay over time. A similar approach is taken in the models presented by Wedel (2004) and Harrington et al. (2018). If the token fails an evaluation, it is not stored. Averaged over many runs, this is equivalent to storing all exemplars with a strength determined by their discriminability and typicality probabilities.

S1.3 Varying discriminability threshold

In the investigation in Sections 5.4–5.5 of the paper, we set discriminability threshold (δ) to be a linear function of type frequency (f):

$$\delta(f) = \left[\lambda + \left(\frac{2(f-1)}{M-1} - 1\right)\phi\right]_0^1 \tag{S13}$$

where M is a constant representing the maximum type frequency in the system (here $M = 12$) and where $[x]_0^1$ evaluates to 0 if $x < 0$, 1 if $x > 1$, and x otherwise. We set a ceiling at $\delta = 1$ because $\delta > 1$ would imply a *dis*advantage for real words in the recognition of phonetically ambiguous stimuli (contra Ganong, 1980). We set a floor at $\delta = 0$ because it represents the limit case where tokens pass the discriminability threshold regardless of the activations they incite.

We varied the parameter λ in Equation (S13) across 3 values, corresponding to the original (constant) values of δ given in Table S3: for parameter sets (1)-(6), we set $\lambda = 0.25$; for parameter

[3] A reviewer asks about overlapping categories that are naturally skewed, such as voiced and voiceless word-initial stops in English (along the VOT dimension). Our model at present is designed to prevent excessive skewness, so it would not be appropriate for such situations, but future work could look at extensions. For example, skewness could be promoted by introducing external articulatory forces that are asymmetric with respect to acoustic value, such as a bias against prevoicing. Nevertheless, under the assumption that naturally skewed categories do not become *more* skewed as they change, squeezing toward the mode is still superior to squeezing toward the mean.

sets (7)-(12), we set $\lambda = 0.50$; and for parameter sets (13)-(18), we set $\lambda = 0.75$. In each case, we varied the parameter ϕ across 5 values, ranging from 0 to 1.0 in steps of 0.25. This yielded the 15 discriminability functions shown in Figure 9A of the paper, each of which was applied to the corresponding group of 6 parameter sets from Table S3.

S1.3.1 Mechanisms that could derive discriminability asymmetries

By setting the discriminability threshold to vary with type frequency in this way, we introduce an assumption that high-frequency types pass the discriminability threshold more easily than low-frequency types. While this assumption is justified by results in the literature (discussed in Section 5.3 of the paper), its implementation – directly varying the discriminability threshold, δ, with type frequency – does not follow from anything else within the exemplar-based framework. In this section, we outline two theoretically-justified mechanisms from which the assumption could emerge, and we discuss their implications for frequency-based asymmetries in the typicality evaluation.

Under the first mechanism, when an incoming token is perceived, the activation of exemplars is weighted by their structural compatibility with the token (their similarity in phonological frame identity) in addition to their position within the activation window (their similarity in acoustic quality of the target phoneme).[4] Such weighting represents a recognition of the fact that the exemplar space is multidimensional, with dimensions corresponding to the phonological frame as well as the quality of the target category realization (Pierrehumbert, 2002). Thus, the token "map" would activate an exemplar of the type *map* with a given F1 value more than an exemplar of the type *pat* with the same F1 value; exemplars of *map* would receive an activational boost from their high structural compatibility with the token "map" (proportional to their position within the activation window). Because a high-frequency type is represented by more exemplars than a low-frequency type, its category receives more of these activational boosts than it would in an equivalent situation with a low-frequency type, yielding greater expected category activation for high-frequency types than for low-frequency types. The ratio of category activations is thus expected to be greater for a high-frequency type than for a low-frequency type, making it easier to pass the discriminability evaluation. A consequence of this mechanism is that the greater expected category activation for high-frequency types also makes them more likely to pass the typicality evaluation.

Under the second mechanism, high- and low-frequency types project activation windows of different sizes. In defining the exemplar-based Generalized Context Model, Nosofsky (1986, p. 41) states that the perceptual sensitivity parameter[5], c, (inversely related to our activation window size parameter, α) "would be expected to increase... as subjects gained experience with the stimuli". Thus, the perception of a token of a high-frequency type is expected to draw on fewer exemplars that are far from the token in the perceptual-acoustic space than the perception of a token of a low-frequency type. Consequently, the activations of both the intended and the other category are expected to be lower for a token a high-frequency type than for an equivalent token of a low-frequency type; in particular, the activation of the other category is expected to be very small for a token of a high-frequency type relative to a token of a low-frequency type, since most exemplars

[4] Exemplars may also be weighted by their contextual similarity with the token more generally. Weighting according to the broad context provided by talker or situation may generate perceptual adaptation effects, where the listener rapidly adjusts perceptual expectations and representations while listening (Norris et al., 2003; Kraljic & Samuel, 2006; Bradlow & Bent, 2008; Clarke-Davidson et al., 2008; Dahan et al., 2008).

[5] The sensitivity parameter is assumed by Nosofsky (1986) to be constant across all types experienced by a given subject, but it could plausibly be extended to vary across types, given that exemplar-based models assume that experience is accrued in a type-specific manner (Pierrehumbert, 2002). Indeed, Nosofsky (1991) considers the equivalent of such an extension and finds that it gives superior description of human recognition data (in the visual mode), though not of classification data.

of the other category are located far from the average token of the intended category. Thus, the ratio of activations (intended/other) would generally be greater for high-frequency types than for low-frequency types, making it easier to pass the discriminability evaluation. A consequence of this approach is that the lower expected category activation for high-frequency types would make them less likely to pass the typicality evaluation.

What do we make of the different implications for asymmetries in the typicality evaluation? In Section S3.1, we discuss results from the literature suggesting that tokens of high-frequency types are stored in memory less robustly than tokens of low-frequency types, and how that can be incorporated in our model as an assumption that tokens of high-frequency types are less likely to pass the typicality evaluation than tokens of low-frequency types. We also present results of additional simulations showing that such an asymmetry in typicality evaluation reinforces the effects of the asymmetry in discriminability evaluation that we have discussed in Section 5.5 of the paper. These results are consistent with the second mechanism we have presented here (frequency-sensitive activation windows) and highlight how mechanistic assumptions have the potential to unify different perceptual effects.

At present, we have not incorporated either of the mechanisms presented here into our model, but we believe that doing so (either separately or together) could be fruitful for future research. However, we caution that such incorporation will require a great deal of care, since the different mechanisms have different implications for asymmetries in the typicality evaluation, and potentially more broadly. Ultimately, both mechanisms may be relevant, and future research will have to determine how they can work together without interfering with each other.

S2 Parameter tuning

S2.1 Tuning for single-category movement: approach

We illustrate our approach to parameter tuning for the single-category case in Figure S2. The general strategy was to pre-determine values for the initial category width, σ, and then choose values for the other parameters so as to obtain category movement with maintenance of category shape and width.

We pre-determined three values for σ, representing narrow ($\sigma = 0.6$), medium-width ($\sigma = 0.8$), and wide ($\sigma = 1.0$) distributions. We fixed the activation window size, α, to be half the width of the category, σ, reflecting the observation that perception should draw on neither too many nor too few exemplars within a category. Thus, for $\sigma = 0.6$, we set $\alpha = 0.3$; for $\sigma = 0.8$, we set $\alpha = 0.4$; and for $\sigma = 1.0$, we set $\alpha = 0.5$. Fixing α in this way is not problematic because its role is to provide a perceptual scale, moderating the effect of the typicality threshold, τ, on the typicality force.

The goal of parameter tuning is to balance the three forces described in Sections 2.4 and 4.1 of the paper: the intrusive bias force, the spreading imprecision force, and the squeezing typicality force. Balancing the forces in this way will allow us to identify regimes that meet the desiderata of generating category movement while maintaining category shape and width. This can be accomplished by adjusting two forces while keeping the other one fixed (providing the scale), because what matters for the qualitative dynamics of the system is the size of each force relative to the others. We therefore pre-determined three values for the bias size, β, yielding three different (fixed) strengths of the bias force: weak ($\beta = 0.05$), medium-strength ($\beta = 0.15$), and strong ($\beta = 0.25$).

These decisions gave us 9 sets of partially-established parameter values (one for each combination of σ, and β), with two parameters to tune: τ, which determines the strength of the typicality force; and ι, which determines the strength of the imprecision force. We tuned these parameters

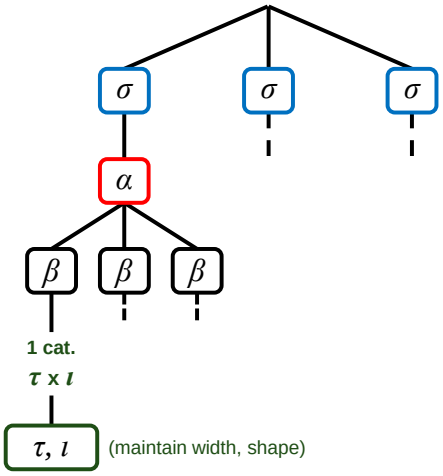

Figure S2: Illustration of the parameter-tuning process for single-category movement. The initialization parameter σ (blue box) was set to several values to define the objectives of the modeling process. The activation window size parameter α (red box) was arbitrarily fixed to $\sigma/2$ to provide a scale (without loss of generality). The bias size β (black boxes) was set to pre-defined (controlled) values. The other relevant parameters, typicality threshold τ and imprecision degree ι (green boxes) were tuned in order to meet the desiderata of maintaining category width and shape.

by varying them independently among 10 values each, with τ ranging from 0.02 to 0.20 in steps of 0.02 and ι ranging from 0.1 to 1.0 in steps of 0.1. The smallest value of τ represented a requirement for the activation incited by a token to be 2% of the maximum possible in order to be stored with probability 0.5, and the largest value represented a requirement for the activation incited by a token to be 20% of the maximum possible in order to be stored with probability 0.5. The smallest value of ι represented a degree of imprecision that could shift the target by up to 6% of the span of a category in either direction (in a wide-category parameter combination), and the largest value represented a degree of imprecision that could shift the target by up to the entire span of a category (in a narrow-category parameter combination).

For each of the 100 pairs of values of τ and ι and each of the nine pairs of values of σ (and corresponding value of α) and β, we ran the model 100 times for 5000 iterations (enough to indicate the equilibrium state). For each value of σ, we chose a value of τ that resulted in a stable shape of the exemplar distribution which was minimally different to the initial shape (i.e. had a minimal increase in kurtosis) and allowed the category to shrink or grow depending on ι; in each case, we chose $\tau = 0.10$. Then, for each value of pair of values of σ and τ, we chose the values of ι that best maintained category width and shape (skewness and kurtosis). For $\sigma = 0.6$, we chose $\iota = 0.3$; for $\sigma = 0.8$, we chose $\iota = 0.4$; and for $\sigma = 1.0$, we chose $\iota = 0.5$.

S2.2 Tuning for single-category movement: results

We summarize the results of the single-category tuning in Table S2 (for initial category properties, see Table S1).

The tuning process allowed us to confirm that parameter choice had the expected implications for category properties. For example, category displacement grew approximately linearly with bias (β), independent of category width (initial value σ), imprecision degree (ι), activation window size (α), and typicality threshold (τ). Increasing bias also increased category skewness, as exemplar movement under bias became more extreme relative to what might be expected under imprecision.

Table S2: Tuned parameter values and average category properties for a single category after 5000 iterations. Displacement measures the distance traveled by the category centroid.

Parameters					Category properties			
σ	β	ι	α	τ	Displacement	Width	Skew	Ex. Kurtosis
0.6	0.05	0.3	0.3	0.10	0.35	0.59	−0.12	−0.48
0.8	0.05	0.4	0.4	0.10	0.35	0.79	−0.09	−0.46
1.0	0.05	0.5	0.5	0.10	0.32	0.97	−0.07	−0.49
0.6	0.15	0.3	0.3	0.10	1.03	0.61	−0.33	−0.20
0.8	0.15	0.4	0.4	0.10	1.06	0.79	−0.23	−0.34
1.0	0.15	0.5	0.5	0.10	1.04	0.99	−0.21	−0.37
0.6	0.25	0.3	0.3	0.10	1.68	0.64	−0.47	0.21
0.8	0.25	0.4	0.4	0.10	1.69	0.83	−0.35	−0.08
1.0	0.25	0.5	0.5	0.10	1.72	1.01	−0.28	−0.26

A corresponding effect was observed for category excess kurtosis: as the category became skewed under high bias, it also became more dispersed.

The tuning process also allowed us to identify relationships between parameter values that are necessary for meeting our desiderata. For example, both ι and τ are required to be sufficiently large relative to β in order to prevent the category becoming excessively skewed under the application of production bias. When ι is small relative to β, little of the variation in production can be attributed to imprecision, meaning that the effect of bias is very clear. When τ is small relative to β, few extreme tokens are discarded for being atypical, meaning that bias is permitted to continue unchecked in the creation of extreme tokens. It is a consequence of this result that, for given values of ι and τ, category skewness increases with β. In addition, an increase in ι requires a concomitant increase in τ in order to maintain category shape, and vice-versa. When ι is too large relative to τ, the category becomes wider, and vice-versa when it is too small. This result reflects the careful balance required between the imprecision and typicality forces.

S2.3 Tuning for two-category interaction: approach

Our approach to parameter tuning for two-category interaction was very similar to the approach for single-category movement. The general strategy was to pre-determine values for the initial category width, σ, and initial category distance, μ, and then choose values for the other parameters so as to obtain interactions exhibiting maintenance of category shape, width, distance, and overlap, building on the existing results for single-category movement. We illustrate our approach to parameter tuning for two-category interaction in Figure S3.

We used the same three pre-determined values for σ as in the single-category case, representing narrow ($\sigma = 0.6$), medium-width ($\sigma = 0.8$), and wide ($\sigma = 1.0$) distributions. We also fixed α in the same way as in the single-category case, setting $\alpha = 0.3$ for $\sigma = 0.6$, $\alpha = 0.4$ for $\sigma = 0.8$, and $\alpha = 0.5$ for $\sigma = 1.0$. We pre-determined two values for μ for each value of σ, representing two

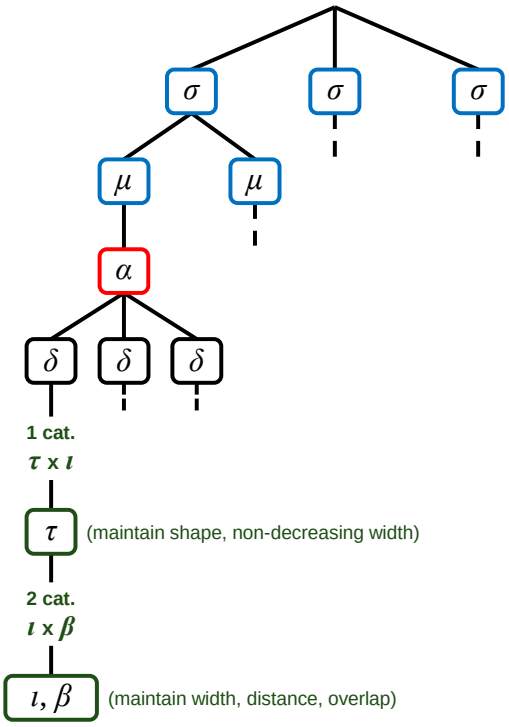

Figure S3: Illustration of the parameter-tuning process for two-category interaction. The initialization parameters σ and μ (blue boxes) were set to several values to define the objectives of the modeling process. The activation window size parameter α (red box) was arbitrarily fixed to $\sigma/2$ to provide a scale (without loss of generality). The discriminability threshold δ (black boxes) was set to pre-defined (controlled) values. The other parameters (τ, ι, β; green boxes) were tuned in order to meet the objectives of the modeling process, in two steps. In the first step, we drew on our simulations of a single category with a range of values of ι and τ to choose a value of τ yielding maintenance of category shape alongside a range of non-decreasing category widths (for different values of ι). In the second step, we simulated two interacting categories with a range of values of ι and β and chose the value of ι that yielded best maintenance of category width and distance and the value of β that additionally yielded best maintenance of category overlap.

different category distances and degrees of category overlap.[6] For $\sigma = 0.6$, we set $\mu \in \{2.1, 1.9\}$; for $\sigma = 0.8$, we set $\mu \in \{3.0, 2.8\}$; and for $\sigma = 1.0$, we set $\mu \in \{3.9, 3.7\}$. While these initialization parameters contribute to the initial behavior of the model – causing, for example, greater initial discriminability force when the overlapping region is initially dense – they have little impact on the long-term dynamics of the model. They thus help to identify cases where the objective of the model has been met, as opposed to affecting the processes that allow this objective to be met.[7]

As in the single-category case, the goal of parameter tuning is to balance forces; to the three forces from the single-category case, the two-category case adds the repulsive discriminability force. To balance four forces, one can be fixed while the others are adjusted. We pre-determined three values for the discriminability threshold, δ, yielding three different (fixed) strengths of the discriminability force: weak ($\delta = 0.25$), medium-strength ($\delta = 0.50$), and strong ($\delta = 0.75$). We did this in order to control discriminability across parameter combinations, so that we could explore the effect of manipulating it consistently with type frequency later (see Section S1.3). The three values of δ we chose were all less than 1, yielding higher discriminability (of intended types) than would be expected based on activations alone; this follows the fact that phonetically ambiguous tokens are biased towards being recognized as real words rather than non-words (Ganong, 1980).

These decisions gave us 18 sets of partially-established parameter values (one for each combination of σ, μ, and δ), with three parameters to tune: β, which determines the strength of the bias force; ι, which determines the strength of the imprecision force; and τ, which determines the strength of the typicality force.

We drew upon our previous single-category simulations to choose a suitable value for τ and a suitable range of values for ι, thus identifying potential typicality and imprecision forces. We retained the value of τ that allowed for best maintenance of category shape ($\tau = 0.10$). Then, for each pair of values of σ and τ, we chose 4 values of ι which yielded a range of category widths, from no increase over time to an increase of up to 50% over time. For $\sigma = 0.6$, we chose $\iota \in \{0.4, 0.5, 0.6, 0.7\}$; for $\sigma = 0.8$, we chose $\iota \in \{0.5, 0.6, 0.7, 0.8\}$; and for $\sigma = 1.0$, we chose $\iota \in \{0.6, 0.7, 0.8, 0.9\}$. We chose such a range for ι because we reasoned that the addition of the discriminability force in the move to a two-category system was likely to favor additional narrowing of categories, which we needed to counter in order to meet our desiderata.

Finally, we used two-category simulations for each parameter set to narrow down to a single value of ι and identify a suitable value of β, thus completing the balancing of forces. For each value of σ, we considered the 4 values of ι obtained from the single-category tuning process alongside 25 values of β. The values of β ranged from 0.01 to 0.25 in steps of 0.01. The smallest value of β represented a consistent bias approximately equal to 0.2% of the span of a category (in a wide-category parameter combination), and the largest value represented a bias approximately equal to 8.5% of the total span of a category (in a narrow-category parameter combination).

For each of the 100 pairs of values of β and ι, each of the 3 values of δ, and each of the 6 pairs of values of σ and μ (and corresponding value of α), we ran the model 100 times for 5000

[6] The values of μ were chosen to yield the same span of the overlapping region in absolute terms (i.e. 0.6 and 0.8 units), regardless of the category width. This means that, for narrower categories, a larger proportion of the category distribution was located in the overlapping region.

[7] There is nothing special about our pre-determination of σ and μ; we do not intend to suggest that these are the *only* values that are appropriate for the model. *Every* combination of the other parameters yields some equilibrium behavior of the system, many of which (i.e. those in which the shape of category distributions does not greatly change) correspond to some combination of σ and μ. However, not every such equilibrium behavior will meet our model desiderata, both in the sense of corresponding to the behaviors we are attempting to model (e.g. some will reflect mutual repulsion of categories) and in the sense of demonstrating appropriate properties (e.g. some will not permit a large degree of category overlap). We chose specific values of σ and μ that would allow us to meet the model desiderata, but there are many other values which would also do so.

Table S3: Tuned parameter values for two-category interaction.

Set	σ	μ	β	ι	α	δ	τ
(1)	0.6	2.1	0.08	0.5	0.3	0.25	0.10
(2)	0.6	1.9	0.10	0.5	0.3	0.25	0.10
(3)	0.8	3.0	0.06	0.6	0.4	0.25	0.10
(4)	0.8	2.8	0.08	0.6	0.4	0.25	0.10
(5)	1.0	3.9	0.05	0.7	0.5	0.25	0.10
(6)	1.0	3.7	0.07	0.7	0.5	0.25	0.10
(7)	0.6	2.1	0.12	0.5	0.3	0.50	0.10
(8)	0.6	1.9	0.16	0.5	0.3	0.50	0.10
(9)	0.8	3.0	0.08	0.6	0.4	0.50	0.10
(10)	0.8	2.8	0.12	0.6	0.4	0.50	0.10
(11)	1.0	3.9	0.12	0.8	0.5	0.50	0.10
(12)	1.0	3.7	0.15	0.8	0.5	0.50	0.10
(13)	0.6	2.1	0.20	0.6	0.3	0.75	0.10
(14)	0.6	1.9	0.25	0.6	0.3	0.75	0.10
(15)	0.8	3.0	0.16	0.7	0.4	0.75	0.10
(16)	0.8	2.8	0.21	0.7	0.4	0.75	0.10
(17)	1.0	3.9	0.14	0.8	0.5	0.75	0.10
(18)	1.0	3.7	0.19	0.8	0.5	0.75	0.10

iterations (enough to indicate the equilibrium state). For each set of values of σ, μ, and δ, we chose the value of ι and a set of values of β yielding best maintenance of category width and distance (allowing slight increases in width and decreases in distance if strict maintenance fell outside the range of parameter values considered). From the set of values of β thus obtained, we chose the one that yielded best maintenance of category overlap (assessed in terms of the area between the two category distributions; this consistently resulted in maintenance or increase of the total span of the overlapping region).

S2.4 Tuning for two-category interaction: results

As described above, we selected one value of each of τ, β, and ι for each set of values of σ, μ, and δ (and corresponding value of α), so as to obtain maintenance of category shape, width, distance, and overlap. This gave 18 sets of parameter values that yielded suitable category interactions. These parameter values are given in Table S3. The average properties of the interactions obtained under these sets of parameter values (after 5000 iterations) are given in Table S4.

Table S4: Average properties of the interactions obtained under the sets of parameter values in Table S3 after 5000 iterations. Overlap measures the span of the overlapping region between categories (i.e. the distance between the most advanced Pusher exemplar and the least advanced Pushee exemplar). Pushee displacement measures the distance traveled by the Pushee centroid (i.e. the size of the push).

			Pushee				Pusher		
Set	Category dist.	Overlap	Displacement	Width	Skew	Ex. Kurtosis	Width	Skew	Ex. Kurtosis
(1)	2.09	0.80	0.08	0.62	0.12	−0.56	0.61	−0.17	−0.54
(2)	1.91	0.96	0.11	0.61	0.11	−0.54	0.60	−0.18	−0.52
(3)	3.01	0.85	0.07	0.81	0.10	−0.57	0.81	−0.12	−0.56
(4)	2.81	1.00	0.10	0.80	0.11	−0.55	0.80	−0.15	−0.55
(5)	3.91	0.88	0.05	1.01	0.10	−0.57	1.00	−0.09	−0.57
(6)	3.69	1.03	0.07	1.00	0.11	−0.57	0.99	−0.12	−0.56
(7)	2.08	0.68	0.12	0.60	0.15	−0.55	0.58	−0.23	−0.50
(8)	1.90	0.83	0.17	0.59	0.14	−0.53	0.57	−0.26	−0.45
(9)	3.02	0.67	0.11	0.80	0.13	−0.57	0.79	−0.17	−0.56
(10)	2.80	0.82	0.14	0.78	0.14	−0.55	0.77	−0.21	−0.51
(11)	3.87	0.80	0.11	1.02	0.14	−0.57	1.00	−0.18	−0.55
(12)	3.68	0.95	0.15	1.01	0.14	−0.57	0.99	−0.21	−0.53
(13)	2.10	0.67	0.18	0.62	0.17	−0.54	0.59	−0.30	−0.45
(14)	1.94	0.83	0.23	0.61	0.15	−0.51	0.57	−0.32	−0.39
(15)	2.99	0.68	0.17	0.81	0.16	−0.55	0.79	−0.25	−0.50
(16)	2.79	0.83	0.21	0.80	0.16	−0.54	0.77	−0.28	−0.46
(17)	3.89	0.67	0.14	1.00	0.14	−0.58	0.99	−0.21	−0.54
(18)	3.67	0.81	0.19	0.99	0.15	−0.57	0.96	−0.25	−0.50

As in the single-category case, the tuning process revealed that certain relationships between parameters were required in order to meet our desiderata. For example, the imprecision force was required to be substantially large (ι near σ in value) in order to maintain category width. It is imprecision that allows targets to be produced outside of the existing exemplar distribution for a given category, facilitating the retreat of the Pushee. Such facilitation is only possible if the imprecision force is large enough to overcome the typicality force. This conclusion is reinforced by the fact that the value of ι chosen in the tuning process for two-category interactions was consistently larger than the value chosen in the tuning process for single-category movement.

Similarly, greater bias force (higher values of β) was required to maintain smaller category distances (smaller values of μ). It is the bias force that causes the Pusher to move toward the Pushee, countering the repulsion due to the discriminability force, and nearer categories have denser overlapping regions and thus greater discriminability force. Consistent with this observation, greater bias force was also required when the discriminability threshold was higher (higher values of δ), yielding greater discriminability force.

Finally, the tuning process highlighted the generality of the model's ability to meet our desiderata. The tuning process began with arbitrary decisions of values for σ, μ, α, and δ. That these decisions did not limit our ability to identify sets of parameter values that allowed us to meet our desiderata suggests that there are many such suitable sets of parameter values. Furthermore, the model was not highly sensitive to the particular value of some tuned parameters. For example, slightly larger values of τ and ι would have yielded similar category shapes and widths to the values chosen, and thus would also have been appropriate.

S3 Additional simulations

In various parts of the paper, we indicated a number of additional kinds of simulations that we had conducted to back up claims about the model's dynamics. In this section, we present the details and results of these simulations.

S3.1 Varying the typicality threshold

In the literature, there is ample evidence that tokens of low-frequency words are generally stored in memory more robustly than tokens of high-frequency words, and consequently have greater potential to impact the perceptual system post-exposure. Relative to high-frequency words, low-frequency words: are more easily recognized as repeated or non-repeated (Schulman, 1967); are rated as more memorable (Benjamin, 2003); benefit more from prior study in identification tasks (Wagenmakers et al., 2000); yield more repetition priming (Forster & Davis, 1984), which spans modalities (Bowers, 2000) and operates regardless of other attentional demands (Kinoshita, 1995); inhibit the recognition of phonetically similar words more in phonetic priming (Goldinger et al., 1989); elicit greater activation in memory-sensitive brain regions when studied (Chee et al., 2004); and may attract attention more, yielding larger attentional blinks and less disruption from attentional blinks (Wierda et al., 2013), and requiring more processing resources to be encoded episodically (Diana & Reder, 2006). In this section, we show that encoding this perceptual asymmetry in the typicality evaluation – by setting the typicality threshold, τ, to be an increasing function of type frequency – causes high-frequency types to change faster than low-frequency types in the Pusher and slower than low-frequency types in the Pushee, reinforcing the effects of varying the discriminability threshold with type frequency.

To introduce frequency-based asymmetries in typicality, we set the typicality threshold (τ) to

be a linearly increasing function of type frequency (f):

$$\tau(f) = \left[\kappa - \left(\frac{2(f-1)}{M-1} - 1\right)\psi\right]^{+} \tag{S14}$$

where, as in the definition of δ functions (Equation (S13)), M is a constant representing the maximum type frequency in the system (here $M = 12$) and where $[x]^{+}$ evaluates to 0 if $x < 0$ and to x otherwise. We set a floor at $\tau = 0$ because it represents the limit case where tokens are never discarded for being atypical, and thus are stored whenever they pass the discriminability threshold. We refer to κ as a measure of the average acceptability of atypical tokens; as κ increases, atypical tokens are less likely to be accepted overall on average. We refer to ψ as a measure of the frequency-based asymmetry in the acceptability of atypical tokens; as ψ increases, tokens of low-frequency types are more likely to be accepted relative to equivalent tokens of high-frequency types.

We varied κ across 3 values: 0.1 (the original value of τ used in all simulations in the paper), 0.2, and 0.3. We varied ψ across 5 values: 0 (no asymmetry, as in all simulations in the paper), 0.05, 0.1, 0.2, and 0.3. Together, this gave us 15 τ functions, illustrated in Figure S4A. We conducted simulations applying these τ functions to 2 parameter sets with different δ functions from our original runs in Section 5 of the paper: one with δ defined by $\lambda = 0.25$ and $\phi = 0.25$, and one with δ defined by $\lambda = 0.75$ and $\phi = 0$ (the other parameters in each case were taken from the corresponding entries in Table S3 with $\sigma = 0.8$ and $\mu = 2.8$, i.e. parameter sets (4) and (16) respectively).

In Section 5.5 of the paper, $(\lambda, \phi) = (0.25, 0.25)$ generated an expected frequency effect (low-frequency type advantage in the Pushee), while $(\lambda, \phi) = (0.75, 0)$ generated a reversed frequency effect. We show that treating τ as a function of type frequency removes the reversed frequency effects. For each τ function and each parameter set, we ran the model 1000 times for 5000 iterations each. The results of the simulations are shown in Figure S4B.

As can be seen, with even a small asymmetry in the typicality threshold favoring the acceptance of atypical tokens of low-frequency types, the reversed frequency effects observed under some parameter settings in Section 5.5 of the paper (i.e. with insufficient asymmetries in discriminability evaluation) disappear. The introduction of typicality asymmetries also strengthens frequency effects that stem from discrimability asymmetries.

We note that the reversed frequency effects are also removed simply by increasing τ (i.e. decreasing the acceptability of atypical tokens) in the absence of frequency-based asymmetries. This result follows – perhaps counterintuitively – from the fact that high-frequency types are more sensitive to perceptual forces than low-frequency types, as a consequence of our assumptions about production and storage (see Appendix B of the paper). Increasing τ increases the squeezing typicality force. Since high-frequency types are more sensitive to this increase, the high-frequency sub-distribution is squeezed more than the low-frequency sub-distribution, causing it to evacuate the overlapping region between categories more. Being further from the overlapping region, the high-frequency sub-distribution is subjected to a lower repulsive discriminability force – but, due to its high sensitivity to this force, it experiences a push of a similar size to that experienced by the low-frequency sub-distribution. The counteracting of discriminability and typicality in this way removes the reversed frequency effect. In the same way, increasing τ more generally causes discriminability-based frequency effects to become less pronounced.

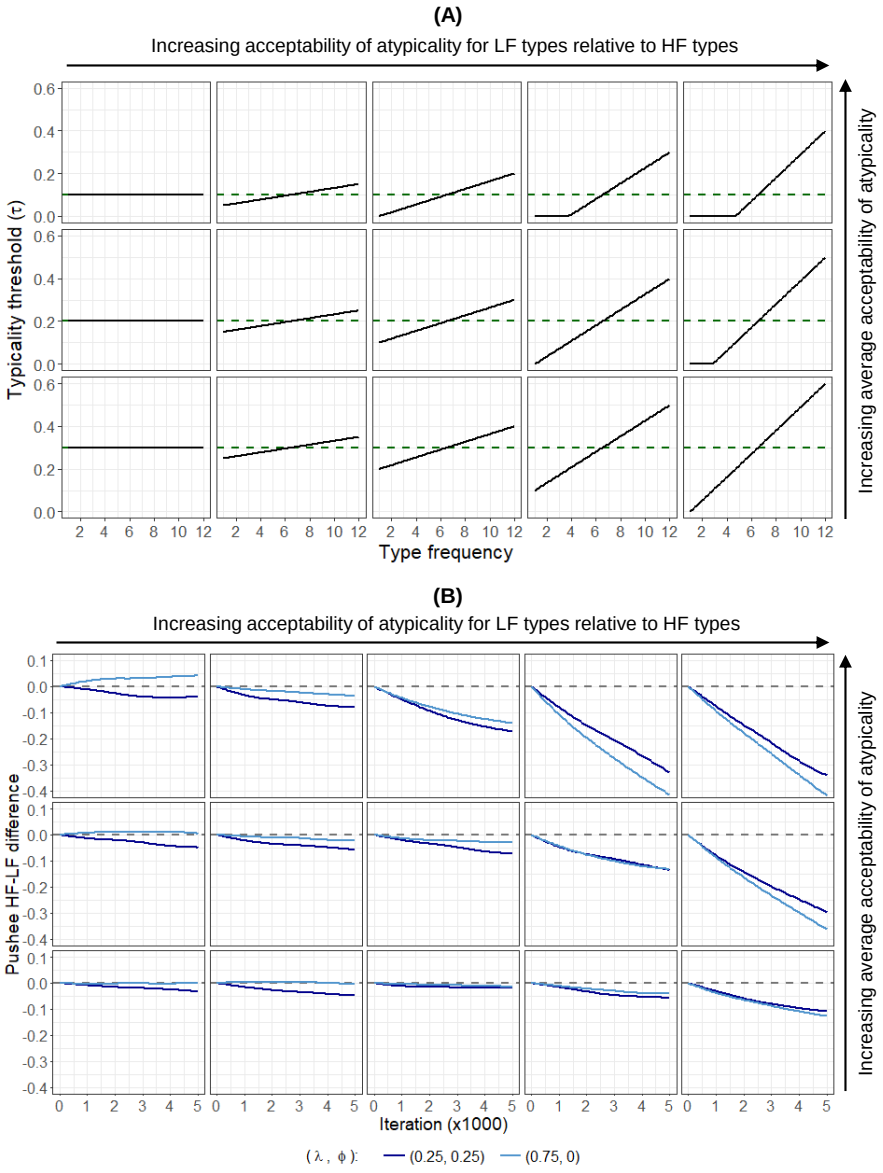

Figure S4: Details and results of treating typicality threshold (τ) as a function of type frequency. (A) τ functions investigated (black lines). Lower τ indicates greater acceptability of atypical tokens. Across all panels in a given row, τ is kept constant for median-frequency types (dashed green lines). This median-frequency τ decreases moving up the rows, making acceptability of atypical tokens higher on average. Across all panels in a given column, the difference between τ for low-frequency types and τ for high-frequency types (slope) is kept constant. This difference increases (slope steepens) moving rightward across the columns, making low-frequency types increasingly more acceptable when atypical than high-frequency types. (B) Results of varying typicality threshold (τ) with type frequency for 2 different sets of parameter values. The vertical axis shows the extent to which high-frequency types are ahead of low-frequency types in the Pushee, averaged over 1000 runs for each parameter setting (blue curve). A positive slope represents a faster rate of change of high-frequency types compared to low-frequency types. As in (A), panels are laid out according to τ function. Moving rightward across the columns, low-frequency types become increasingly more acceptable when atypical than high-frequency types. This shifts the end of the curve downward, causing negative-sloping sections where high-frequency types change at a slower rate than low-frequency types. This effect is present for all choices of average acceptability of atypical tokens.

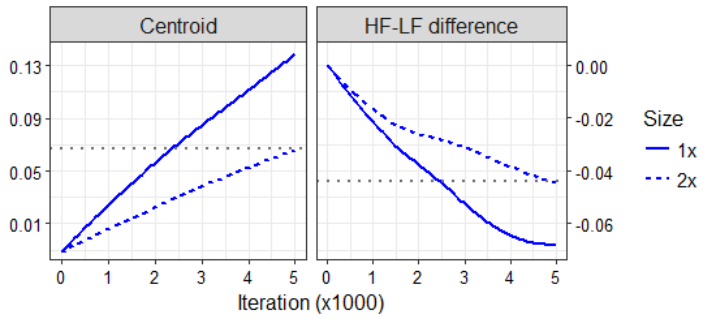

Figure S5: Results of simulations illustrating how the size of the system (solid: 492 exemplars, as in all simulations in the paper; dashed: 984 exemplars, 2x the amount in all simulations in the paper) affects its rate of evolution. For both the centroid (left) and the frequency effect (right) of the Pushee, the system from the paper evolves at approximately twice the rate of a system with twice as many exemplars, i.e. it takes only half as many iterations to reach the same value.

S3.2 Increasing the number of types

In Section 3.1.2 (footnote 8) of the paper, we stated that the rate of evolution of the system is approximately inversely proportional to the number of exemplars it contains. Here, we present the results of simulations that further support the statement; we show that doubling the number of types (and hence, the number of exemplars) causes the system to evolve about half as quickly. We provide further mathematical discussion of the relationship between rate of evolution and number of exemplars in Section S5.2.3.

We created a system with twice as many exemplars by duplicating the contents of each category in the initialization file. We ran 1000 models for 5000 iterations each, using parameter set (10) from Table S3 and a value of 0.50 for ϕ, such that δ decreased linearly from a value of 1 for the lowest-frequency types to a value of 0 for the highest-frequency types.

Figure S5 shows the results of the simulations. As can be seen, the system with twice as many exemplars evolved at approximately half the rate, with respect to both the movement of categories (as represented by the Pushee centroid) and the internal organization of categories (as represented by the Pushee frequency effect).

S3.3 Changing bias

In the paper, we focused on the centrality of the listener to sound change. However, the speaker also plays a role. We stated that the primary role of the speaker in the present model is to ensure that category interaction is persistent, while making minimal contributions to frequency effects in this interaction (relative to the listener).

In this section, we present two sets of simulations where we varied the application of production bias in order to support the statement that the speaker is not as important as the listener for generating frequency effects in the model. We first present simulations where both categories receive production bias, in which the results show that the existence of frequency effects is not tied to overall category movement. We then present simulations where neither category receives productions bias, in which the results show that frequency effects are also not tied to bias and can be obtained even in the absence of speaker influence. In both cases, we continue to refer to the categories as "Pushee" and "Pusher" to allow comparison with our other results. These names should only be taken as convenient, not as indicating the kind of movement exhibited by the category or the bias to which it is submitted.

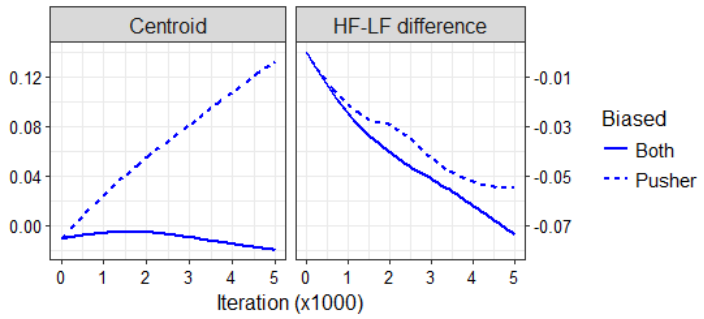

Figure S6: Results of simulations involving two categories biased together (solid line) or a single Pusher category biased toward the other (dashed line), showing the centroid (left panel) and the frequency effect (right panel) of the Pushee. While the centroid hardly moves in the case with two categories biased together, a robust frequency effect is still observed, which is at least as large as the effect observed when just the Pusher is biased.

S3.3.1 Categories biased together

To demonstrate that the model's results on frequency effects stem from the internal reorganization of categories due to perceptual asymmetries, rather than from the movement of categories due to the application of production biases, we conducted simulations in which the two categories were biased together (to the same extent). To accomplish this, we subjected the Pusher to a positive bias that was half the size of that in Section 5 of the paper and the Pushee to a negative bias of the same magnitude. We ran 1000 models for 5000 iterations each, using parameter set (10) from Table S3 and a value of 0.50 for ϕ, such that δ decreased linearly from a value of 1 for the lowest-frequency types to a value of 0 for the highest-frequency types.

In these simulations, both the Pushee and the Pusher stayed approximately still (with slight movement of the centroids due to reversion to the modes, which were slightly off-centered in the initialization data). The results for the Pushee are shown in Figure S6. As can be seen, while biasing the categories together results in almost no overall movement, it still yields a Pushee frequency effect: high-frequency types in the Pushee become peripheral slower than low-frequency types. This frequency effect is at least as large as the effect observed when just the Pusher is biased.

The fact that a frequency effect is observed even without category movement implies that the frequency effects predicted by the model are not dependent on category movement. Rather, frequency effects arise as a result of internal reorganization of categories to balance forces from production and perception, as discussed in Section 5.5 of the paper. The fact that the same kind of frequency effect is obtained under two qualitatively different kinds of production force implies that it is driven by the perceptual forces, i.e. by processes in the listener rather than the speaker.

S3.3.2 No bias

Having established that the model's results on frequency effects are independent of category movement due to production bias, we conducted further simulations to demonstrate that they are independent of production bias altogether. To accomplish this, we removed the bias entirely, such that neither category was biased in any way. We ran 1000 models for 5000 iterations each, using parameter set (10) from Table S3 – with the exception that $\beta = 0$ – and a value of 0.50 for ϕ, as before.

In these simulations, the Pushee and the Pusher were repelled from one another, and gradually drifted apart. This repulsion was caused by the discriminability force, which caused perceptual downweighting of tokens produced near the region of category overlap. Its gradualness was a result

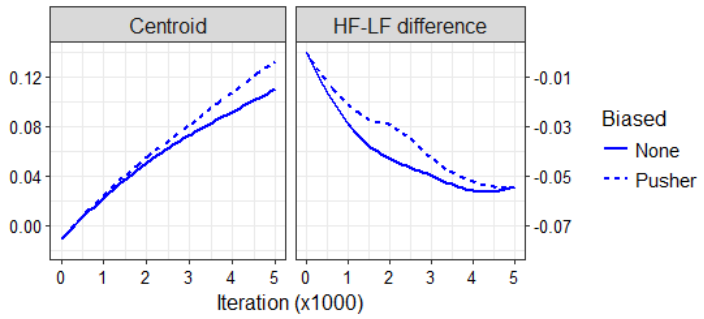

Figure S7: Results of simulations involving two categories without bias (solid line) or a single Pusher category biased toward the other (dashed line), showing the centroid (left panel) and the frequency effect (right panel) of the Pushee. While the removal of bias causes decreased centroid movement, it does not affect the frequency effect.

of the typicality force, which caused similar downweighting of tokens produced far from the mode of each category. Taken together, these two forces yield the *hyperspace effect* (Johnson et al., 1993): from the listener's perspective, the optimal production target for a category is hyperarticulated – i.e. located further away from other categories than the mode – but not so much as to no longer resemble natural speech.

The results for the Pushee are shown in Figure S7. As can be seen, the degree of repulsion decreases in the absence of Pusher bias as the two categories separate, yielding less category movement over time. However, the frequency effect appears to be unaffected: high-frequency types change slower than low-frequency types in the Pushee, regardless of whether there is production bias in the system or not. The fact that frequency effects are still observed even when production bias is removed confirms that such results in the model follow from the listener, not the speaker. The role of the speaker is to enable prolonged category interaction, by counteracting the repulsion of the Pusher with production bias – but interaction does not have to be prolonged in this way in order for categories to internally reorganize and display frequency effects.

S3.4 Adding minimal pairs

In the paper, all of our simulations involved the simplifying assumption that there are no minimal pairs in the system. We argued (in Appendix A.2) that minimal pairs could not be driving empirically-observed category movements and frequency effects, since such effects are seen across the lexicon, in which most words do not participate in a relevant minimal pair. Our argument can be broken into two sub-arguments: firstly, that minimal pairs do not make necessary contributions to any part of modeling empirically-observed category movements and frequency effects; and secondly, that minimal pairs alone are not sufficient to generate these movements and effects. In this section, we present an extension of the model to include minimal pairs, together with additional simulations from this extended model, to support our arguments.

S3.4.1 Modeling minimal pairs

There are two options for introducing minimal pairs into the model, differentiated based on the assumed influence of higher-level (syntactic, semantic, pragmatic, or discourse) context. Under the first, context-sensitive option, higher-level context has a large influence: it uniquely determines the intended type even when the phonological frame is consistent with multiple types. The listener operates as in the present model, storing the token as an exemplar of the intended type if it passes the discriminability and typicality evaluations, and discarding it otherwise. Consequently, there is

no potential for *variant trading* (Blevins & Wedel, 2009), where the listener mistakenly stores a token of one type as an exemplar of another type. Under the second, context-insensitive option, higher-level context has no influence: when the phonological frame is consistent with multiple types, the context can never uniquely determine which type was intended. The listener considers all possible types that are consistent with the phonological frame, including those corresponding to nonwords, and chooses one probabilistically based on category activation and type frequency. The token is stored as an exemplar of the winning type if it passes the typicality evaluation and if the type corresponds to a real word; thus, variant trading is possible.

Within the framework of the model, the context-sensitive option is more conservative because it does not permit variant trading; in every other respect, the two options are mathematically equivalent. Without variant trading, all that introducing minimal pairs does is effectively raise the discriminability threshold. Recall that the threshold was stated to be low ($\delta < 1$) due to the existence of lexical bias (Ganong, 1980) toward the intended type, making it relatively easy to map an acoustically ambiguous token to the intended type. With the introduction of minimal pairs, the unintended type can also have such a lexical bias, introducing another plausible identity for an acoustically ambiguous token, and thus making it harder to map such a token to the intended type. As shown mathematically in Equation (S17), this counterveiling pressure effectively raises δ. Since we have already explored the role of different discriminability thresholds in Sections 5.4–5.5 of the paper without invoking minimal pairs, introducing minimal pairs under the context-sensitive option would not yield any new insight. For this reason, we chose to introduce minimal pairs under the context-insensitive option, allowing us to explore anew the influence of variant trading on the model's results.

The introduction of context-insensitive minimal pairs embraces the view of the discriminability evaluation as a probabilistic recognition process involving competition between the types (from different categories) that are consistent with a given frame (see Section S5.1). We assume that a token is identified as belonging to the category that wins the discriminability evaluation, and that the discriminability evaluation fails just in case this identification yields a nonword. More precisely, the discriminability evaluation (Equation (S10)) is replaced with a process of *recognition* of the token (Equation (S15)), where a δ value is computed as before for each category according to the corresponding type frequency (Equation (S16)). The token is passed to the typicality evaluation if and only if it has been recognized as corresponding to a real word.

$$P(\text{token recognized as } T_k) = \frac{\frac{1}{\delta_k} \cdot A_k}{\sum_k \frac{1}{\delta_k} \cdot A_k} \tag{S15}$$

$$\delta_k = \begin{cases} \left[\lambda + \left(\frac{2(f_k - 1)}{M - 1} - 1\right)\phi\right]_0^1 & T_k \text{ corresponds to a real word} \\ 1 & T_k \text{ corresponds to a nonword} \end{cases} \tag{S16}$$

When two competing types both correspond to real words, the recognition equation can be explicitly written out as follows:

$$P(\text{token recognized as } T_i) = \frac{\frac{1}{\delta_i} \cdot A_i}{\frac{1}{\delta_i} \cdot A_i + \frac{1}{\delta_o} \cdot A_o} \tag{S17}$$

$$= \frac{\frac{\delta_o}{\delta_i} \cdot A_i}{\frac{\delta_o}{\delta_i} \cdot A_i + 1 \cdot A_o} \tag{S18}$$

$$= \frac{\frac{1}{\delta_i / \delta_o} \cdot A_i}{\frac{1}{\delta_i / \delta_o} \cdot A_i + 1 \cdot A_o} \tag{S19}$$

It can be seen that this form is equivalent to Equation (S10), with $\delta = \delta_i/\delta_o$. Since $\delta_o < 1$, it follows that this new version of δ is increased from the original value of δ_i it would take were there no real word competitor, and hence that the introduction of minimal pairs effectively raises the discriminability threshold.

S3.4.2 Minimal pairs are not necessary: simulations with a subset of minimal pairs

We begin by questioning whether minimal pairs are necessary for generating any desirable pattern in simulations of two-category activations. We compare simulations with and without minimal pairs to see whether the addition of minimal pairs makes new results possible or existing results impossible.

In each model run of our simulations, we randomly changed 10% of the types in the system to participate in minimal pairs, as an approximation to the proportion of minimal pairs in the /æ/-/ɛ/ interaction in New Zealand English (see Appendix A.2 the paper). In each run of the model, we randomly chose 10 types from each category and created minimal pair relations between them. This random pairing process removed any influence of type frequency on the results.

The addition of minimal pairs allows categories to stably exist in closer proximity to one another, since tokens that would otherwise fail the discriminability evaluation instead participate in variant trading. To ensure that this decreased stable distance between categories did not disrupt our interpretation of the simulation results, we retuned the initial category distance (μ) using frequency-insensitive discriminability thresholds (δ). Keeping all other parameters as in Table S3, we obtained the following: for $\sigma = 0.6$, we set $\mu \in \{1.8, 2.0\}$; for $\sigma = 0.8$, we set $\mu \in \{2.6, 2.8\}$; and for $\sigma = 1.0$, we set $\mu \in \{3.4, 3.6\}$.

For each of the parameter settings in Table S3 (with μ retuned as above), we ran 1000 models for 5000 iterations each. The average properties of the interactions obtained in models with 10% minimal pairs under these sets of parameter values (after 5000 iterations) are given in Table S5. As can be seen, the properties in Table S5 are highly similar to those in Table S4, indicating that the same kinds of stable category interactions are generated with and without minimal pairs.

It is possible that the similarities between the simulations with and without minimal pairs hold only at the coarse-grained category level and not at the fine-grained type level. To assess this possibility, we repeated the investigation of frequency effects from Sections 5.4–5.5 of the paper in systems involving minimal pairs. Using the 15 δ functions from Figure 9A of the paper and the 18 sets of parameter values with retuned μ, we ran the model 1000 times for 5000 iterations each. In Figure S8, we compare frequency effects in the Pushee for models with and without minimal pairs for parameter sets (4), (10), and (16) (with μ retuned as above).

The system with minimal pairs shows the same broad patterns in frequency effects as the system without minimal pairs: when high-frequency types are sufficiently perceptually advantaged relative to low-frequency types (with respect to discriminability, δ), they become more likely to cluster in the overlapping region between categories, allowing low-frequency types in the Pushee to change at a faster rate. Wherever a robust frequency effect of this sort exists in the system without minimal pairs, it also exists in the system with minimal pairs.

However, the addition of minimal pairs also exacerbates the existence of reversed frequency effects for some δ functions (lower-left panels of Figure S8), where high-frequency types in the Pushee change at a faster rate than low-frequency types. As discussed in Appendix B of the paper, these reversed frequency effects are the result of an interaction between our assumptions about production and storage, which causes high-frequency types to be more sensitive to perceptual forces than low-frequency types. Because the addition of minimal pairs allows for categories to

Table S5: Average properties of the interactions obtained in models with 10% minimal pairs under the sets of parameter values in Table S3 (with μ retuned) after 5000 iterations. Overlap measures the span of the overlapping region between categories (i.e. the distance between the most advanced Pusher exemplar and the least advanced Pushee exemplar). Pushee displacement measures the distance traveled by the Pushee centroid (i.e. the size of the push).

				Pushee				Pusher		
Set	μ	Category dist.	Overlap	Displacement	Width	Skew	Ex. Kurtosis	Width	Skew	Ex. Kurtosis
(1)	2.0	1.99	0.89	0.08	0.62	0.12	−0.55	0.61	−0.18	−0.53
(2)	1.8	1.80	1.07	0.11	0.61	0.12	−0.53	0.59	−0.19	−0.49
(3)	2.8	2.82	1.02	0.07	0.81	0.11	−0.56	0.80	−0.14	−0.56
(4)	2.6	2.62	1.16	0.09	0.80	0.12	−0.55	0.79	−0.16	−0.53
(5)	3.6	3.63	1.12	0.06	1.00	0.11	−0.57	1.00	−0.11	−0.57
(6)	3.4	3.42	1.28	0.08	0.99	0.12	−0.56	0.98	−0.13	−0.55
(7)	2.0	1.99	0.77	0.12	0.60	0.15	−0.54	0.58	−0.25	−0.49
(8)	1.8	1.80	0.94	0.16	0.59	0.13	−0.52	0.57	−0.27	−0.42
(9)	2.8	2.85	0.85	0.11	0.79	0.14	−0.56	0.78	−0.18	−0.54
(10)	2.6	2.64	1.01	0.15	0.78	0.14	−0.54	0.76	−0.22	−0.50
(11)	3.6	3.62	1.04	0.13	1.01	0.14	−0.56	0.99	−0.19	−0.54
(12)	3.4	3.43	1.18	0.17	0.99	0.14	−0.55	0.97	−0.22	−0.52
(13)	2.0	2.01	0.77	0.18	0.61	0.17	−0.54	0.59	−0.31	−0.43
(14)	1.8	1.85	0.94	0.23	0.60	0.15	−0.49	0.57	−0.32	−0.37
(15)	2.8	2.83	0.85	0.17	0.81	0.16	−0.56	0.78	−0.26	−0.49
(16)	2.6	2.63	1.01	0.22	0.79	0.16	−0.52	0.76	−0.28	−0.45
(17)	3.6	3.64	0.94	0.16	1.00	0.15	−0.56	0.97	−0.22	−0.52
(18)	3.4	3.43	1.10	0.22	0.98	0.15	−0.55	0.95	−0.25	−0.48

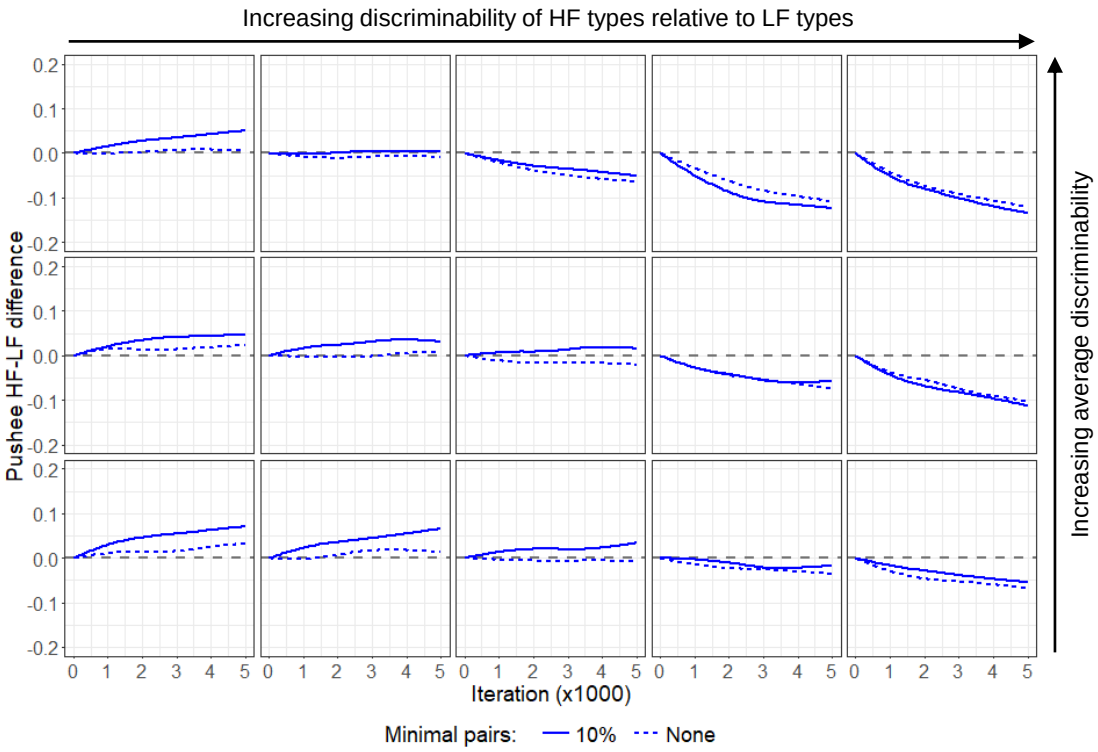

Figure S8: Results of varying discriminability threshold (δ; see Section S1.3) with type frequency for 3 different sets of parameter values (1 per row), comparing a system with 10% minimal pairs (solid lines) to a system with no minimal pairs (dotted lines). The figure is laid out in the same way as Figure 9 of the paper. The system with minimal pairs shows the same patterns as the system without minimal pairs, with increasing discriminability of high-frequency types relative to low-frequency types (movement from left to right across columns) causing slower change of high-frequency types than of low-frequency types in the Pushee (negative-sloping sections). However, it also shows reversed effects when there is little or no difference in discriminability between high- and low-frequency types (left columns), unlike the system without minimal pairs.

stably exist closer to one another, it increases the discriminability force,[8] High-frequency types are more sensitive to this increased discriminability force, causing them to be pushed apart moreso than low-frequency types in the absence of a countervailing perceptual asymmetry. We consider the size of the reversed frequency effect not to qualify as a meaningful difference between the models with and without minimal pairs, since the effect is an artifact of our simplified assumptions about production and storage, and since it only occurs in situations where an empirically-supported perceptual asymmetry is not present.

The addition of minimal pairs thus does not meaningfully affect the model's results. Models both with and without minimal pairs are equally capable of generating two-category interactions displaying key properties observed in documented sound changes, such as the maintenance of category width and overlap. Furthermore, given sufficiently strong perceptual asymmetries, models both with and without minimal pairs generate word-frequency effects of the kind observed in documented sound changes. We conclude that minimal pairs are not necessary for generating any desirable pattern in simulations of two-category activations.

S3.4.3 Minimal pairs are not sufficient: simulations with only minimal pairs competing

Given that the model's key results can be obtained both with and without minimal pairs, we next ask whether they can be obtained if minimal pairs alone underpin category interaction. We conduct simulations varying the degree to which types with and without minimal partners contribute to category interaction, to see whether minimal pairs alone are sufficient for generating desirable category movements and frequency effects.

Since we assume that the phonological frame is perfectly perceived (Section S1.2.6), we assume that recognition of a type involves competition only between types with the same phonological frame, i.e. between two real words in a minimal pair or between a real word and a nonword. Given this assumption, to say that minimal pairs alone underpin category interaction is to say that types corresponding to nonwords do not compete with types corresponding to real words for recognition. This lack of nonword type competition is a tacit assumption in existing models of spoken word understanding (e.g. Norris & McQueen, 2008).[9] In the model presented in the paper (described in Section 3; simulated in Section 5), we assume that nonword types compete to the degree that would be expected based on their category activation alone, by setting the default value of δ for nonwords to 1. Here, we relax this assumption by increasing the default value of δ; the larger the value, the less nonword types compete, and thus the more minimal pairs carry the burden of category interaction.

To control the extent to which nonword types compete with real words during the recognition process, we introduced a new parameter, χ. χ is a scale factor that multiplies the activation of a category for the purpose of recognition when the corresponding type is a nonword, just as $\frac{1}{\delta}$

[8] To see why decreased category distance results in increased discriminability force, consider a token at the edge of the intended category, in the overlapping region. The discriminability force is a function of the number of exemplars of the other category contained within the activation window around this token; more exemplars from the other category provide more competition during the discriminability evaluation, yielding a larger discriminability force. The closer the categories are, the closer the edge of the intended category will be to the centroid of the other category, and thus the more exemplars from the other category there will be in the activation window.

[9] Models of spoken word understanding typically assume that competition is between real words that may be phonological neighbors without being minimal pairs in regards to the segment in question (vowel); for example, *bat* competes not just with *bet*, but also with words like *pat* and *back*. This assumption is a consequence of the phonological frame not being perfectly perceived, and would also follow in our model if we allowed imperfect frame perception. However, extending the model in this way is beyond the scope of the present work, and it is not clear that it would systematically contribute to the interaction between vowel categories.

multiplies the activation when the corresponding type is a real word (Equation (S20)). In this way, $\frac{1}{\chi}$ corresponds to the default δ value assigned to nonword types. χ can be interpreted as (proportional to) the response bias toward a category yielding a nonword type.

$$\delta_k = \begin{cases} \left[\lambda + \left(\frac{2(f_k-1)}{M-1} - 1\right)\phi\right]_0^1 & T_k \text{ corresponds to a real word} \\ \frac{1}{\chi} & T_k \text{ corresponds to a nonword} \end{cases} \tag{S20}$$

When the unintended ("other") type corresponds to a nonword, the formula underlying recognition as the intended type (Equation (S15) can be written out explicitly as:

$$P(\text{token recognized as } T_i | T_o \text{ corresponds to a nonword}) = \frac{\frac{1}{\delta_i} \cdot A_i}{\frac{1}{\delta_i} \cdot A_i + \chi \cdot A_o} \tag{S21}$$

When $\chi = 0$, the right-hand side of Equation (S21) becomes 1, meaning that every type that is not in a minimal pair relation is automatically correctly recognized (because there is only one real word compatible with the perfectly-perceived phonological frame). In other words, nonword types do not compete for recognition. When $\chi = 1$, the recognition process reverts to that explored in the previous subsection, in which nonword types compete for recognition to the same extent as in Section 5 of the paper, but trigger failure when they win. For intermediate values of χ, nonword types have intermediate degrees of influence on the recognition process.

Note that Equation (S20) is equivalent to Equation (S22), where $\delta'_k = \delta_k \chi$. Consequently, introducing the parameter χ is equivalent to multiplying both λ and ϕ by a scale factor. In other words, reducing the extent to which nonword types compete with real word types for recognition is equivalent to increasing the average discriminability of types (lowering the discriminability threshold) and decreasing the discriminability of high-frequency types relative to low-frequency types.

$$\delta'_k = \begin{cases} \left[\lambda\chi + \left(\frac{2(f_k-1)}{M-1} - 1\right)\phi\chi\right]_0^1 & T_k \text{ corresponds to a real word} \\ 1 & T_k \text{ corresponds to a nonword} \end{cases} \tag{S22}$$

To test how χ affects category interaction and type frequency effects, we conducted simulations. Our simulations involved 10% minimal pairs, as in Section S3.4.2, and used parameter setting (10) from Section S3.4.2, with a value of 0.5 for ϕ. To ensure that we could focus just on category interaction, independent of external effects, we removed Pusher bias by setting $\beta = 0$. We explored 6 values for χ: 0, 0.1, 0.25, 0.5, 0.75, and 1. For each value of χ, we ran 1000 models for 50000 iterations each. We present a summary of the rest of the average results after 50000 iterations in Table S6. As can be seen, category shape (width, skewness, and excess kurtosis) is approximately maintained for all values of χ, but increases of category distance and Pushee frequency effects are only obtained for $\chi > 0$. We present a summary of how category distance and Pushee frequency effects change over time in Figure S9.

It is clear from Table S6 and Figure S9 that having an extremely large default value of δ – corresponding to no recognition competition from nonword types – is not appropriate, for two main reasons. Firstly, the categories drift closer together over time to greatly increase overlap, in spite of the expectation that they should be mutually repellent. Secondly, types of all frequencies change at the same rate in the Pushee, in spite of the expectation that perceptual asymmetries should allow low-frequency types to change faster (as in the previous versions of the model). Both of these results follow from the fact that, when nonword types do not compete for recognition, an intended type without a minimal partner is automatically recognized regardless of its frequency, because it is the only real word type that is compatible with the perfectly-perceived phonological frame.

Table S6: Average values and % changes for properties of interactions after 50000 iterations, for models with 10% minimal pairs where nonword types compete during recognition to various degrees (represented by χ). The models use parameter set in Table S3, retuned to have $\mu = 2.6$, and have no bias ($\beta = 0$). Pushee frequency effect measures the distance between the centroid of the sub-distribution of high-frequency Pushee types and the centroid of the sub-distribution of low-frequency Pushee types.

					Pushee						
χ	Category distance		Category overlap		Width		Skew		Ex. Kurtosis		Freq. Effect
0	2.45	(− 5.8%)	2.26	(+125.8%)	0.90	(+11.3%)	0.02	(−79.9%)	−0.49	(+34.2%)	0.00
0.1	2.97	(+14.3%)	1.36	(+ 36.2%)	0.87	(+ 7.8%)	0.07	(−15.5%)	−0.53	(+28.1%)	−0.03
0.25	3.26	(+25.3%)	0.95	(− 5.2%)	0.86	(+ 6.6%)	0.07	(− 5.8%)	−0.55	(+25.4%)	−0.03
0.5	3.47	(+33.5%)	0.65	(− 34.6%)	0.86	(+ 6.3%)	0.09	(+13.5%)	−0.57	(+23.7%)	−0.03
0.75	3.60	(+38.3%)	0.50	(− 49.8%)	0.86	(+ 5.9%)	0.09	(+16.0%)	−0.57	(+23.6%)	−0.05
1	3.69	(+42.1%)	0.40	(− 59.7%)	0.86	(+ 5.8%)	0.09	(+16.1%)	−0.56	(+24.1%)	−0.04

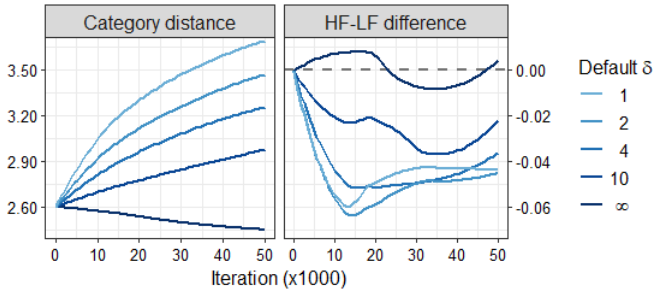

Figure S9: Results of simulations involving two categories with no bias, where the default value of δ for nonword types is increased to represent less nonword type competition in recognition, from the standard value of 1 (lightest; expected competition) to the largest value possible (darkest; no competition). Left: the distance between the categories grows in all cases except when there is no nonword type competition. Right: high-frequency types change slower than low-frequency types in all cases except when there is no nonword type competition.

Thus, there is no discriminability force for the 90% of types without minimal partners, meaning that there is insufficient force to keep the categories apart, and there is insufficient potential for perceptual asymmetries to be leveraged in the generation of frequency effects.

Conversely, any default δ that is not extremely large – i.e. any non-zero degree of competition from nonword types – is sufficient to generate mutual category repulsion and frequency effects. While smaller default δ (more nonword type competition) causes greater increase in category distance, it has little impact on category shape, nor on the degree to which low-frequency types change faster than high-frequency types in the Pushee. Consequently, the model reported in the paper – where nonword types compete fully, i.e. as would be expected based on the category activations they incite – yields a qualitative pattern of results that is expected to hold even if the degree of nonword type competition is reduced.

In summary, the model's key results cannot be obtained if minimal pairs alone underpin category interaction (at least, assuming that only a minority of types are in relevant minimal pair relations, as indicated by the New Zealand English corpus data discussed in Appendix B of the paper). The burden for category interaction must be extended to types without minimal partners, so that phonotactically plausible nonword types compete for recognition (even to a small degree). Since it is types without minimal partners that are crucial to the model's key results, we conclude that the decision to leave out minimal pairs from the model in the paper had no qualitative effect on our main results (reported in Section 5.5).

S3.4.4 How minimal pairs contribute: All types as minimal pairs

The exclusion of minimal pairs in the paper did not cause or prevent any particular kind of category dynamics or frequency effects, given the assumption that categories interact in more ways than just through minimal pairs. However, the addition of minimal pairs appeared to exacerbate reversed type frequency effects that exist in the absence of perceptual asymmetries. In this section, we test the degree to which the addition of minimal pairs affects frequency effects, by running simulations where every type is in a minimal pair relation.

In each run of the model, we randomly paired each type with a type in the other category. As before, this random pairing process removed any influence of type frequency on the results.

To ensure that the decreased stable distance between categories did not disrupt our interpretation of the simulation results, we again retuned the initial category distance (μ) using frequency-insensitive discriminability thresholds (δ). Since our previous investigations suggested that minimal pair presence did not interact with category width, we only performed simulations with the three parameter sets with $\sigma = 0.8$ that we used to compare frequency effects with and without minimal pairs in the previous subsections: sets (4), (10), and (16). For these parameter settings, our retuning process led us to set $\mu = 1.4$. Using these parameter settings, we ran 1000 models for 5000 iterations each. We illustrate the results for different δ functions in Figure S10, comparing a system made up of all minimal pairs to a system with no minimal pairs.

As would be expected based on the results with 10% minimal pairs, giving every word a minimal partner caused large reversed frequency effects in the absence of sufficiently strong perceptual asymmetries. The reason for these large reversed frequency effects is the same as in the 10% case: with every type in a minimal pair, categories can stably exist at very close distances, yielding a very large discriminability force. High-frequency types are more sensitive to perceptual forces than low-frequency types, and thus are repelled more by this very large discriminability force. The resultant effect is larger in the case where every type is in a minimal pair than in the case where 10% of types are in a minimal pair because the categories can stably exist at closer proximities, yielding a larger discriminability force and hence a larger difference between low- and high-frequency types.

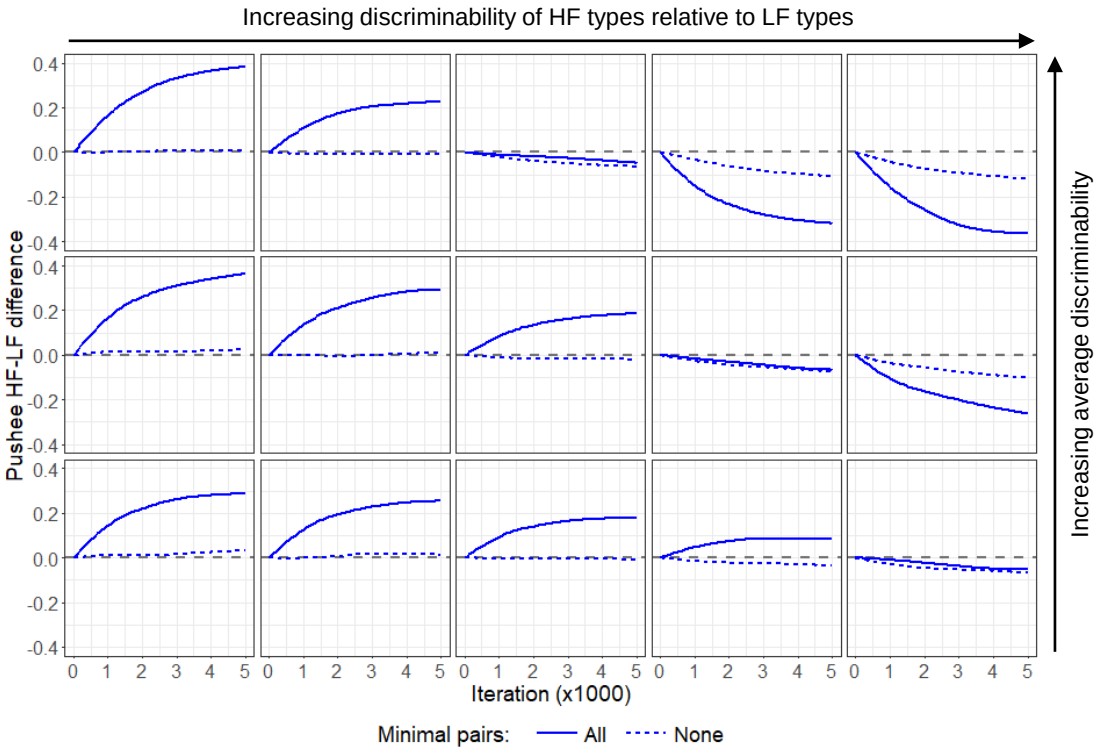

Figure S10: Results of varying discriminability threshold (δ; see Section S1.2.7 for details) with type frequency for 3 different sets of parameter values (1 per row), comparing a system made up of all minimal pairs (solid lines) to a system with no minimal pairs (dotted lines). The figure is laid out in the same way as Figure 9 of the paper. The system with minimal pairs shows the same patterns as the system without minimal pairs, to an extreme degree.

What is unexpected based on the results with 10% minimal pairs, however, is that the frequency effects in the presence of strong perceptual asymmetries – where low-frequency types change faster than high-frequency types – are likewise enlarged relative to the original model. This result highlights an additional implication of making every type a member of a minimal pair, which is that every type can be misrecognized as one from the other category; in other words, every type can participate in *variant trading*. When a range of high-frequency types have sufficiently high discriminability (i.e. δ sufficiently close to 0), they will (practically) always win any recognition process involving their frames, regardless of the intended category. For example, in the top-right panel of Figure S10 (in which the underlying δ function has $\lambda = 0.25$ and $\phi = 0.75$), all types with frequency 8 or above – which includes *all* high-frequency types – have $\delta = 0$, and thus automatically win any recognition process in which they are involved. Consequently, high-frequency types of a given category will be sampled evenly from *both* categories, while low-frequency types will be sampled primarily from their own category. High-frequency types will thus tend toward the mode of the combination of the two category distributions, while low-frequency types will tend toward the mode of their corresponding category distribution. Since the mode of the two distributions combined is between the modes of either distribution independently, it follows that high-frequency types will be over-represented in the overlapping region, while low-frequency types will be over-represented in the non-overlapping (exterior) regions of the categories. While this result is not problematic, the reason for it is, as it implies that a word with a high-frequency minimal partner would never be recognized. Such behavior cautions against the use of systems with too many minimal pairs.

Even in the system composed entirely of minimal pairs, however, there are discriminability functions (δ) that yield entirely reasonable frequency effects, as shown in the three panels in Figure S10 where the system with all minimal pairs overlaps entirely with the system with no minimal pairs. The discriminability functions yielding these results are such that there is a sufficiently large perceptual asymmetry (for the given average discriminability) which does not cause high-frequency types to automatically win recognition. In other words, there are a range of discriminability functions that find a perfect balance for perceptual (discriminability) advantages for high-frequency types – not too little, not too large – under which the exact same frequency effects are generated, regardless of the number of minimal pairs in the system. Either side of these perfectly-balanced discriminability functions, the inclusion of minimal pairs causes frequency effects identified in Section 5.5 of the paper to be exaggerated.

S4 Entrenchment

Our simulations in the paper excluded the process of *entrenchment*, where production is based on an average of exemplars rather than a single exemplar (Pierrehumbert, 2001). We made this exclusion because, at a high level, entrenchment appears to be redundant and inappropriate in comparison to our proposed typicality evaluation. More specifically: (1) a primary motivation of entrenchment is to provide a squeezing force, but it is not the only way to do so – a squeezing force is also provided by our treatment of typicality in perception, which is not based on averaging exemplars; (2) entrenchment squeezes each category toward the mean, which is not as good at preventing excessive skewness of overlapping categories as the squeezing toward the mode due to typicality; (3) given the production-perception loop, averaging of exemplars (if it occurs) need not be based in production – building on Hintzman (1986), Goldinger (1998) argues that averaging occurs in perception, with activated exemplars yielding an *echo* that forms the basis of further processing including storage; (4) under the interpretation of our model as representing an aggregate over a community, it is not clear that averaging exemplars in production is appropriate – a speaker may

not be as influenced by exemplars of others' speech in production as she is by her own.

Nevertheless, since entrenchment is a standard process, we have included an implementation of it in the code accompanying the paper, for future research. In this section, we describe the entrenchment process and the technical details of our implementation.

S4.1 Description

Entrenchment represents production influences stemming internally from the exemplar system, such as practice effects. The process of entrenchment makes the target more similar to existing exemplars of the same category. All of the exemplars of the given category in an entrenchment window around the target are activated, with exemplars near the target activated more than those far away. The target then attempts to shift toward each exemplar in the window, by an amount proportional to that exemplar's activation. The net result of this is that the target shifts in the direction of greater local exemplar density (i.e. toward the category centroid), by an amount related to the asymmetry in local exemplar density.

The extent to which a target shifts under entrenchment is a function of the entrenchment window size, ϵ. As ϵ increases, the entrenchment window widens to include more exemplars at non-negligible activation levels. Since there are more exemplars located toward the category centroid (relative to the target) than away from it, this means that the local density of exemplars located toward the centroid increases more with ϵ than does the local density of exemplars located away from the centroid. As a result, increasing ϵ highlights the asymmetries in local exemplar density more, causing a larger shift of the target toward the centroid.

At the level of the category distribution, entrenchment yields a squeezing force that squeezes each category toward its centroid, enacting reversion to the mean (see Section S1.2.8 for comparison with reversion to the mode, as enacted by the typicality force). The size of this squeezing force increases with ϵ.

S4.2 Implementation details

In our implementation, entrenchment occurs between target selection and the application of bias.

In entrenchment, the target value v is replaced by a weighted average v' taken over all exemplars x in the target category C_i, as shown in Equation (S23). Exemplars are weighted by a fixed function w_e according to their distance from the target.

$$v' = \frac{\sum_{x \in C_i} x w_e(v - x)}{\sum_{x \in C_i} w_e(v - x)} \tag{S23}$$

The treatment of entrenchment as a weighted average is common to most exemplar-based models of the production-perception loop, though they typically differ on details. For example, the model presented by Wedel (2012) and Wedel & Fatkullin (2017) weights exemplars both by distance from the target and by recency, where recency is encoded in time-decaying exemplar strength. The model presented by Wedel (2006) randomly chooses a small set of exemplars (with probability determined by their distance from the target) and weights only these according to recency; averaged over many runs, this produces equivalent behavior to weighting every exemplar by target-distance and recency. The model presented by Wedel (2004) uses a much coarser notion of recency for weighting, giving weight only to the previous few exemplars. The present treatment shows recency-weighting in the same coarse way; because new exemplars overwrite old exemplars of the same type, only exemplars which are recent enough to be present are given any weight. Averaged over many runs, however, the expected behavior of entrenchment in the present model is the same as that in a

model which weights exemplars by recency in a fine-grained manner. Finally, the model presented by Pierrehumbert (2001, 2002) assigns (recency-based) weights only to the k exemplars nearest the target, effectively meaning that the weighting function w_e is not fixed; rather, a unique weighting function is appealed to for each production.

In the present treatment, the weighting for entrenchment is provided by a Gaussian window w_e with width ϵ (a parameter), as shown in Equation (S24). Exemplars that are very near the target are given weights close to 1, while exemplars that are very far away are given weights close to 0. Increasing ϵ causes exemplars within a wider radius to be given non-negligible weights.

$$w_e(d) = \exp\left(\frac{-d^2}{2\epsilon^2}\right) \tag{S24}$$

A Gaussian entrenchment window is also used in the models presented by Wedel (2004) and Wedel (2006); a common alternative is an exponential window (Wedel, 2012; Wedel & Fatkullin, 2017). Our primary reason for choosing a Gaussian entrenchment window in the present treatment is to mirror the use of a Gaussian activation window (see Section S1.2); the use of the same form of weighting function in both production and perception is parsimonious.

S5 Equivalences to other frameworks

Our model is equivalent to models from other frameworks in various ways. Firstly, the discriminability evaluation in our model can be viewed as an interactive probabilistic recognition process, as in the Bayesian model of Norris & McQueen (2008). Secondly, the expected storage and production behavior of our model, with random overwriting of exemplars, is equivalent to that of a special case of exemplar-based models with exponentially-decaying exemplar strengths (following the framework laid out by Pierrehumbert, 2001), when averaged over many runs. In this section, we mathematically derive these equivalences, and we use them to elucidate the discriminability threshold functions and the rate of evolution of the system respectively.

S5.1 Discriminability as interactive recognition

Discriminability evaluation in our model is framed as a check on the quality of a token that has already been identified. It can alternatively be viewed as an interactive probabilistic recognition process, where the acoustic value of the token is assessed bottom-up for its potential to be a realization of candidate types of different categories, which assert top-down influences. In this interpretation, non-existent types (i.e. nonwords) are considered alongside existing types due to their phonotactic plausibility, but if a non-existent type emerges from the recognition process, it is not stored as an exemplar for future productions.

In the discriminability evaluation, an incoming signal with phonological frame p and target phoneme acoustic value v is evaluated for discriminability as a token of category C_i by comparing the activation of that category, $A_i(v)$ (relative to the activation of the other category), to the discriminability threshold, $\delta_i(p)$. In the absence of minimal pairs, the probability of passing the evaluation, $P_D(C_1, v, p)$, is given by Equation (S10), rewritten below in Equation (S25).

$$P_D(C_1, v, p) = \frac{A_1(v)\frac{1}{\delta_1(p)}}{A_1(v)\frac{1}{\delta_1(p)} + A_2(v)} \tag{S25}$$

The use of type frequency to modulate the discriminability of acoustic signals (via frequency-sensitive threshold $\delta_i(p)$) represents a combination of top-down and bottom-up information, which

is typical in models of word recognition. The mathematical form of discriminability evaluation in the present model is similar to that of Bayesian word recognition (e.g. Norris & McQueen, 2008), represented in Equation (S26)[10] for categories C_i, phonological frame p and acoustic value v.

$$P(C_1|v,p) = \frac{P(v|C_1,p)P(C_1|p)}{P(v|C_1,p)P(C_1|p) + P(v|C_2,p)P(C_2|p)} \tag{S26}$$

We make this similarity apparent by dividing both the numerator and denominator in Equation (S26) by $P(C_2|p)$ and comparing the result (Equation (S27)) to Equation (S25).

$$P(C_1|v,p) = \frac{P(v|C_1,p)\frac{P(C_1|p)}{P(C_2|p)}}{P(v|C_1,p)\frac{P(C_1|p)}{P(C_2|p)} + P(v|C_2,p)} \tag{S27}$$

Taking each $P(v|C_j,p)$ to be approximately proportional[11] to $A_j(v)$, the expressions in Equations (S27) and (S25) become (approximately) identical if a particular relationship holds between the prior and the discriminability threshold:

$$\frac{1}{\delta_1(p)} = \frac{P(C_1|p)}{P(C_2|p)} \tag{S28}$$

i.e.

$$\delta_1(p) = \frac{1 - P(C_1|p)}{P(C_1|p)} \tag{S29}$$

since $P(C_2|p) = 1 - P(C_1|p)$ in a two-category system.

S5.1.1 Using the comparison to understand discriminability thresholds

We can use the relationship between discriminability thresholds in our model and prior probability in the Bayesian model (Equation (S29)) to elucidate the interpretation of discriminability thresholds. For illustration, we plot the relationship for several different discriminability threshold functions in Figure S11. As can be seen, lower discriminability threshold corresponds to higher prior probability, and the mapping between discriminability threshold and prior probability is nonlinear.

This comparison also reveals that the linearly-decreasing discriminability threshold functions explored here yield priors of a reasonable shape. A "reasonable" prior would be one that increases in a decelerating fashion with word frequency, as humans have a tendency to over-estimate the frequency of rare words and under-estimate the probability of common words (Begg, 1974). While Figure S11 shows that the prior corresponding to our choice of discriminability threshold function (orange line) is increasing in an accelerating fashion with type frequency, recall from Section S1.1 that type frequency in our model represents a subjective, flattened (log-transformed) form of actual word frequency. When the horizontal axis in Figure S11 is stretched non-linearly to represent

[10] Following our assumptions about the phonological frame being perfectly retrievable, it can be treated as a condition throughout the equation; this means the prior $P(C|p)$ represents a comparison between two potential members of a minimal pair differing only in target phoneme identity, and the likelihood $P(v|C,p)$ is determined by the acoustic quality of the target phoneme.

[11] $P(v|C_j,p)$ need not be exactly proportional to $A_j(v)$. Both are generated from the exemplar distribution in the same way (via Gaussian convolution; see Section S1.2), under imprecision and activation respectively, but the effects of imprecision and activation will yield different results if the corresponding parameters, ι and α, are different. This was found to be the case in our parameter tuning process, with $\iota > \alpha$ in the two-category interactions we modeled; hence the approximation.

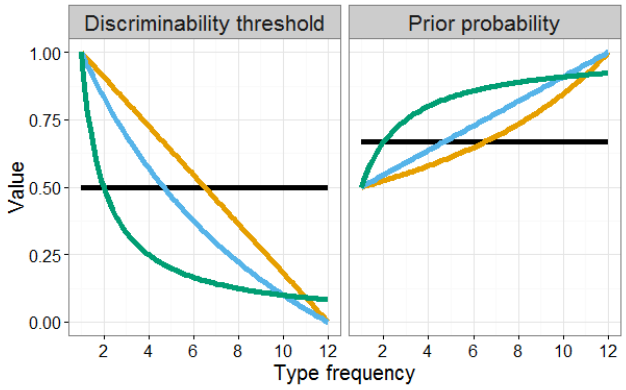

Figure S11: Illustration of the correspondence between discriminability threshold in the present model and prior type probability in a Bayesian framework. Lower discriminability threshold corresponds to higher prior probability. A constant discriminability threshold corresponds to a uniform prior (black). A linearly-decreasing discriminability threshold corresponds to an accelerating increasing prior (orange). As the discriminability threshold becomes more bowed downward, the prior becomes more bowed upward, passing from linearly-increasing (blue) to decelerating increasing (green).

word frequency rather than type frequency, all of the increasing curves are revealed to be decelerating. Thus, a discriminability threshold function that decreases linearly with type frequency corresponds to a prior that increases in a decelerating fashion with word frequency, which we take to be reasonable.

S5.2 Overwriting and decay

In Appendix A.4 of the paper, we asserted that our treatment of memory, where all stored exemplars have constant strength 1 and storage entails overwriting an exemplar of the same type, is equivalent to a treatment in which exemplars decay in strength over time and are never overwritten (following the framework laid out by Pierrehumbert, 2001). Building on this point, we asserted that the expected overall behavior of our "overwriting" model (averaged over many runs) is the same as that of the "decay" model for a particular choice of decay rate, as the expected behavior of the overwriting model (averaged over many runs) is mathematically equivalent to that of a special case of the decay model. Here, we derive these assertions, and we use them to show that the rate of evolution of the system in our model is determined by the total number of exemplars (across types).

For both the overwriting and the decay models, we consider a type with frequency f in a system where the total combined frequency of all types (i.e. total number of exemplars, in the overwriting model) is N. Our mathematical analysis assumes that f represents subjective type frequency. It is not sensitive to the way in which these subjective frequencies are obtained from objective values (e.g. whether they represent raw or log corpus frequencies), though this may have implications for the caveats we describe in Section S5.2.4. We focus on a single exemplar of this type that was stored at time 0, and we consider its contribution to the behavior of the system at time t (i.e. after t production-storage iterations). For simplicity, we assume that every token is stored; allowing some tokens not to be stored slows down the rate of evolution. For the decay model, we assume that the strength of each exemplar is scaled by a factor of $k < 1$ with each iteration. As in the simulations presented in the paper (Section 3.1.5), we also assume that both models consist of a single agent talking to herself, so that the sources of storage and production are identical; our derivations as stated do not apply to situations with multiple interacting agents, which are beyond the scope of

this paper.

Our mathematical derivations assume that there are multiple types in the system, but they make no assumptions about the allocation of those types to categories. Our results can be understood to apply equally well to a case with a single category (and multiple types in that category) or to a case with multiple categories (and at least one type per category). Because we assume there are multiple types, our decay-based model is not the same as the one presented by Pierrehumbert (2001), which observed apparent frequency effects using a single type in a single category. In Section S5.2.5, we develop a full comparison with the actual model presented by Pierrehumbert (2001), through which we demonstrate why the apparent frequency effects observed from that model do not hold of exemplar-based models in general.

S5.2.1 Equivalence of memory treatments

We first show that the overwriting and decay models have equivalent treatments of memory. For this purpose, we compare the probability that an exemplar remains after t iterations in the overwriting model with the strength of an exemplar after t iterations in the decay model.

Overwriting model. In the overwriting model, an exemplar stored at time 0 will remain at time t provided that any subsequent tokens of the same type do not overwrite it.

The probability of producing a token of the given type on any iteration is f/N. Given a token of that type, the probability of overwriting the given exemplar with it is $1/f$. Thus, the probability of overwriting the given exemplar on any iteration is $1/N$, so the probability of *not* overwriting it on any iteration is $1 - (1/N)$.

For the exemplar still to be present after t iterations, it must not have been overwritten on each iteration. Since each iteration is independent, the probability of this is

$$P(\text{exemplar remains at time } t) = \left(1 - \frac{1}{N}\right)^t \tag{S30}$$

which is exponentially decreasing with t at a rate given by $1 - 1/N$.

Decay model. In the decay model, an exemplar is stored at time 0 with strength 1, and this strength decays exponentially.

On each iteration, the strength of the exemplar is multiplied by $k < 1$. Thus, the strength of the exemplar at time t is

$$S_x(t) = k^t \tag{S31}$$

Model comparison. As can be seen, the probability of an exemplar remaining after t iterations in the overwriting model (Equation (S30)) and the strength of an exemplar after t iterations in the decay model (Equation (S31)) have the same form. Furthermore, for the particular choice of $k = 1 - (1/N)$, they are identical. Thus, though the two models appear to have very different treatments of memory, they are mathematically equivalent in terms of their expected outcomes (averaged over many runs).

S5.2.2 Equivalence of overall expected behavior

Having established that the two models have equivalent expected treatments of memory (averaged over many runs), we now show that they have equivalent treatments of production, in terms of their expected choice of production targets (averaged over many runs). Since the system evolves by means of producing new tokens to store in memory, these equivalences jointly imply that the two models are equivalent in terms of their overall expected behavior (averaged over many runs).

We consider an exemplar stored at time 0 and compare the probability of choosing that exemplar as production target at time t in both models.

Overwriting model. In the overwriting model, the choice of a given exemplar as production target at time t has three conditions. Firstly, the exemplar must remain in the system at time t. Secondly, the speaker must choose to produce the type of which the exemplar is an instance. Thirdly, the exemplar must be chosen as target from all exemplars of that type.

The probability of the exemplar remaining in the system at time t is $(1-(1/N))^t$ (Equation (S30)), the probability of the type being chosen is f/N, and the probability of the exemplar being chosen from all f exemplars of that type is $1/f$. Thus, the probability of choosing an exemplar as production target t iterations after it was stored is:

$$P(\text{exemplar chosen as target at } t) = \frac{1}{N}\left(1-\frac{1}{N}\right)^t \tag{S32}$$

Decay model. In the decay model, the choice of a given exemplar as production target at time t has two conditions: the speaker must choose to produce the corresponding type, and the exemplar must be chosen as target from all exemplars of that type.

As in the overwriting model, the probability of the type being chosen is f/N. The probability of choosing the exemplar from all exemplars of that type is $S_x(T)/\sum_{y\in T} S_y(t)$, where $S_x(t) = k^t$ is the strength of the exemplar at time t and $\sum_{y\in T} S_y(t)$ is the total strength of all exemplars of that type at time t. Thus, the probability of choosing an exemplar as production target t iterations after it was stored is:

$$P(\text{exemplar chosen as target at } t) = \frac{f}{N}\cdot\frac{k^t}{\sum_{y\in T} S_y(t)} \tag{S33}$$

For the sake of exploring expected behavior (i.e. behavior on average, over many runs), $\sum_{y\in T} S_y(t)$ may be approximated by $S^\star$, the expected total strength at any time. To obtain a value for $S^\star$, we consider r synchronized runs of the model (where r is large) at a particular point in time, with total strengths $\widehat{S_i^\star}$ (for i from 1 to r) for a given type. $S^\star$ is given by the mean of these total strengths.

$$S^\star = \frac{\sum_{i=1}^{r}\widehat{S_i^\star}}{r} \tag{S34}$$

After a single iteration, each total strength $\widehat{S_i^\star}$ will have been multiplied by k due to decay, and $(f/N)r$ of them are expected to have also grown by 1 due to new productions of the given type. Their mean is still expected to be $S^\star$.

$$S^\star = \frac{\left(\sum_{i=1}^{r} k\widehat{S_i^\star}\right) + \frac{f}{N}r}{r} \tag{S35}$$

$$= k\frac{\sum_{i=1}^{r}\widehat{S_i^\star}}{r} + \frac{f}{N} \tag{S36}$$

Substituting Equation (S34) into Equation (S36) yields

$$S^\star = kS^\star + \frac{f}{N} \tag{S37}$$

which can be solved for $S^\star$:

$$S^\star = \frac{f}{N(1-k)} \tag{S38}$$

Substituting $S^\star$ for $\sum_{y \in T} S_y(t)$ in Equation (S33) gives an analytic approximation of the expected probability (averaged over many runs) of choosing an exemplar as production target t iterations after it was stored:

$$P(\text{exemplar chosen as target at } t) \approx (1-k)k^t \tag{S39}$$

Two caveats are required in order for this approximation to be valid. Firstly, the system must not be in its early iterations. Secondly, the decay rate must not be extremely fast relative to the range of type frequencies, such that low-frequency types are expected to have total exemplar strength $S^\star < 1$. We describe the caveats in more detail in Section S5.2.4.

Model comparison. As can be seen, the probability of an exemplar being chosen as a production target t iterations after it was stored has the same form in both the overwriting (Equation (S32)) and the decay (Equation (S39)) models. As was the case for memory (Section S5.2.1), these probability expressions are identical for the particular choice of $k = 1 - (1/N)$. Thus, the overwriting model's expected overall behavior (averaged over many runs) is a special case of the decay model's expected overall behavior (averaged over many runs). Given an overwriting model with a total number of exemplars N, it is possible to choose a decay rate k allowing the construction of a decay model with the same expected overall behavior (averaged over many runs). Consequently, any overwriting model is equivalent to some decay model.

We note, however, that the reverse equivalence is not always true: for some decay models, it is not possible to construct an overwriting model showing the same expected overall behavior (averaged over many runs). Because an overwriting model necessarily contains at least one exemplar of each type at every point in time, it requires all types to have expected exemplar strength $S^\star \geq 1$, which is not true in decay models in which the decay rate is extremely fast (relative to the range of type frequencies). See Section S5.2.4 for further discussion.

S5.2.3 Using the comparison to understand rate of evolution

Exemplar models evolve via the production and storage of tokens. Assuming (for simplicity) that each production involves the application of a constant bias and all tokens are stored, the rate of evolution of the system – i.e. the rate with which the production bias translates into category movement – is determined by the probability distribution over production targets according to their age. The more likely recent exemplars are to be chosen as targets, the more production biases will snowball, and the faster the system will evolve.

In the overwriting model (Equation (S32)), the probability of an exemplar being chosen as target t iterations after it was stored is uniquely determined by N, the total number of exemplars in the system. For small N, the probability mass is concentrated around small t, so that recent exemplars are much more likely than old exemplars to be chosen as targets and thus the system evolves quickly. As N increases, the probability mass remains highest for small t, but spreads out over a larger range of values of t, increasing the probability that older exemplars will be chosen as targets and thus decreasing the rate of evolution of the system. Thus, increasing the number of exemplars in the system simply slows down its evolution, without affecting its qualitative behavior, as shown by the simulations in Section S3.2. A similar observation can be made for the decay model (Equation (S39)) with respect to the decay parameter k; decreasing k causes the system to evolve faster.

Because the probability distribution over production targets according to their age does not vary with type frequency (i.e. f is not involved in Equations (S32) and (S39)), it follows that types of all frequencies are expected (on average) to evolve at the same rate. This is the basis of the lack of frequency effect observed for single-category movement under the overwriting model (see Section 4.2 of the paper). Under the decay model, a lack of frequency effect is also obtainable (see Sóskuthy, 2014, for related simulations), but it is contingent on the caveats for the assumptions made in our analysis. In Section S5.2.4, we describe the caveats in detail and show that there are certain parameter settings under which the decay model can give rise to frequency effects in single-category movement.

S5.2.4 Caveats for the decay model

In Section S5.2.2, we noted that there are two caveats on the analytical approximation for production target selection in the decay model. Both caveats concern the approximation of the total strength of all exemplars of a particular type T at time t, $\sum_{y \in T} S_y(t)$, by the expected total strength at any time, $S^\star$ (Equation (S38)).

Firstly, the system must "burn in" – i.e. be run for sufficiently many iterations – in order for strengths to build to the expected value $S^\star$. In other words, the exemplar distributions for each type must build up to stable densities before the approximation is valid. Thus, our analytical approximation does not hold for the early iterations of a decay-based model that is seeded from sparse exemplar distributions. Consequently, it would not apply to a situation such as a child accumulating experience as they learn a language. However, given that we model regular sound change, which can occur within a lifetime and be reflected in the way that an adult's speech changes (Harrington, 2006), we do not believe this limitation to prevent us from gaining insight from our model comparison.

Secondly, the system must be defined in such a way that $S^\star$ is sufficiently greater than 1. When $S^\star$ is close to 1, the total strength $\sum_{y \in T} S_y(t)$ is volatile, as the addition of 1 strength with each new exemplar constitutes a substantial portion of $S^\star$. In this case, $\sum_{y \in T} S_y(t)$ will tend to be above $S^\star$ for small t, meaning that the approximation will tend to overestimate the probability of recent exemplars being selected as target. Consequently, types for which $S^\star \lessapprox 1$ will not have recent exemplars selected as production targets as often – and thus will not advance as rapidly – as expected under the approximation. Since $S^\star$ decreases with type frequency (Equation (S38)), extreme cases of the decay model (i.e. ones in which $S^\star \approx 1$ for low-frequency types and $S^\star \gg 1$ for high-frequency types) may thus predict low-frequency types to advance at a slower rate than high-frequency types. Such extreme cases would arise in the presence of either an extremely fast decay rate or an extremely long-tailed distribution of type frequencies, where a high-frequency type is presented for storage orders of magnitude more often than a low-frequency type.

In what follows, we illustrate how choices made by the modeler can affect this second caveat, radically altering the qualitative results of a decay-based model (assuming it has been run for sufficient iterations first, as in the first caveat). To facilitate our illustration, we introduce two quantities of interest: the *e-folding time* for the system and the *recurrence time* for different types. The e-folding time, defined in Equation (S40), is related to the decay rate and represents the number of iterations required for exemplar strength to decay by a factor of e. From the e-folding time, we can also obtain the *exemplar lifespan*, representing the number of iterations for which an exemplar persists in memory; for the following discussion, we assume that an exemplar may be removed once its strength depletes by more than 99% (following Wedel & Fatkullin, 2017), giving a lifespan of approximately 5 e-folding times. The recurrence time, defined in Equation (S42), is the reciprocal of (normalized) type frequency and represents the expected number of iterations between

productions of a given type.

$$e\text{-folding time: } E = \frac{-1}{\ln(k)} \tag{S40}$$

$$\approx \frac{1}{1-k} \tag{S41}$$

$$\text{recurrence time: } R = \frac{N}{f} \tag{S42}$$

Equations (S41)[12] and (S42) can be substituted in Equation (S38) to yield a definition of expected total exemplar strength, $S^\star$, in terms of e-folding time and recurrence time, given in Equation (S43).

$$S^\star \approx \frac{E}{R} \tag{S43}$$

The definition in Equation (S43) allows us to easily recognize when the second caveat will not hold, and thus when frequency effects will be expected. We expect frequency effects in a model in which high-frequency types have recurrence times much shorter than the e-folding time and low-frequency types have recurrence times at least as long as the e-folding time.

How can the e-folding and recurrence times be determined? Both are measured in terms of model iterations. An iteration corresponds to the production and perception of a single token that is considered for storage. Thus, some guidance can be provided by consideration of the objective rates of production and perception of words in the real world. After accounting for sampling error (Pierrehumbert & Granell, 2018), objective recurrence times for different words can be estimated to range from 20 (for the word *the*) to more than 100 million (for extremely rare words). To put these numbers in context, Brysbaert et al. (2016) calculate that the average person may hear just under 12 million words per year, and a typical psycholinguistic study (e.g. Carreiras et al., 2006) defines "high-frequency" words as having recurrence times of approximately 25,000 (40 tokens per million) and "low-frequency" words as having recurrence times between 300,000 and 2 million (3 to 0.5 tokens per million). However, these objective distributions do not translate directly into the model. Since not every actual word token that is uttered need be considered for storage (as discussed in Appendix A.3 of the paper), the representations of type frequency in the model – and thus the determinations of the e-folding and recurrence times – rely on subjective distributions. The modeler is free to choose the function mapping from objective to subjective distributions, giving a large amount of freedom over the choice of e-folding and recurrence times. This freedom of choice can determine model behavior.

For example, if we assume that subjective frequencies are identical to objective frequencies, then the extremely large range of recurrence times means that there is a correspondingly large range of e-folding times in which a model will show frequency effects. For example, any e-folding time around 2 million iterations or less – corresponding to an exemplar lifespan of 10 months or more – will generate lag among "low-frequency" words as defined by the psycholinguistic literature. The literature does not contain enough results on the processing of rare words to determine whether this long exemplar lifespan is appropriate for low-frequency words, but the use of objective frequencies

[12]The approximation in Equation (S41) is obtained from taking the first-order Taylor polynomial of $\ln(k)$ about 1 and holds provided k is sufficiently close to 1. For example, for all cases discussed here ($k > 0.9995$, $E \geq 2000$), the multiplicative error in the estimation is 0.025% or less, which does not substantially impede our ability to identify circumstances in which $S\star \lessapprox 1$ or $S^\star \gg 1$.

implies that it must also apply to high-frequency words, for which it is likely too long. Consequently, any choice of *e*-folding time that is not too long for high-frequency words will cause some types to change faster than others in a model using objective frequencies.

In such a situation, precisely *which* types change faster will depend upon the *e*-folding time. For example, with an *e*-folding time of 2,000 (following Pierrehumbert, 2001), faster change would be observed among types with recurrence times of less than around 2,000. For English, this corresponds to a small set of about 200 extremely common words, which does not have good coverage of the content words defined as "high-frequency" in the prior literature. Consequently, the frequency effects obtained in this situation would not correspond to real effects observed empirically. Under an alternative *e*-folding time of 30,000, the set of faster-moving types would expand to include the approximately 3,000 English words typically defined as "high-frequency". In this situation, an exemplar would have a lifespan of approximately 5 days, which is extremely fast in comparison to the recurrence times for low-frequency words that occur around once a year (or less). Consequently, a model assuming this *e*-folding time and distribution of recurrence times would also have to assume that rare words – which encompass a non-negligible proportion of the lexicon, as demonstrated in Figure S1 – are practically incapable of establishing stable exemplar-based representations in the minds of typical speakers. Such an assumption would raise questions for studies drawing on representations of rare words, such as *mammary* in Bybee's original work adducing a connection between word frequency and leniting changes (Hooper, 1976).

Alternatively, if we assume that subjective frequency is nonlinearly "flattened" from objective frequency, then the range of recurrence times is likewise compressed, and it becomes easier for models to show no frequency effects. For example, in our current model, recurrence times range from 41 (for the highest-frequency type) to 492 (for the lowest-frequency type).[13] In a corresponding decay model with an *e*-folding time of 2,000 iterations (again following Pierrehumbert, 2001), we would thus expect no frequency effects. To provide an indication of what this *e*-folding time means on a real-time scale, we change the interpretation of subjective frequencies. Previously, we interpreted subjective frequencies as reflecting the assumption that some tokens are filtered out in perception before being considered for storage, with more filtering for higher-frequency types. Alternatively, we can maintain that all tokens are considered for storage, and interpret subjective frequencies as reflecting the assumption that exemplars of higher-frequency types are stored with lower initial strength, giving them shorter lifespans. Under this interpretation, an *e*-folding time of 2,000 means that "high-frequency" words (as defined in the psycholinguistic literature) would have lifespans of around 1–2 months, while "low-frequency" words would have lifespans of around 5–10 months. Much work remains to be done in investigating a real-world-scale decay version of our model, but such work goes beyond the scope of the present discussion.

To summarize, the question of whether or not a decay-based model meets the second caveat – and thus whether it displays no frequency effects or a high-frequency advantage – depends on the relationship between the *e*-folding time, determined by the decay rate, and the distribution of recurrence times, determined by subjective type frequencies. To generate frequency effects, the *e*-folding time needs to be sufficiently fast and the range of recurrence times sufficiently large. Furthermore, when there are frequency effects, the subset of words that change faster is determined by where the *e*-folding time falls in the distribution of recurrence times.

[13] It is a consequence of our overwriting-based treatment of memory that the recurrence time for the lowest-frequency type can be no greater than the total number of exemplars in the system. For reasons of computational resources, our present simulations assume a small number of types and hence a small number of exemplars, giving us relatively short recurrence times. Scaling up the number of types in our model will also scale up the recurrence times. If the *e*-folding time is not scaled up commensurately, then there are certain parameter ranges in which we expect to observe frequency effects (where the *e*-folding time falls between the recurrence times for high- and low-frequency types).

The e-folding time and distribution of recurrence times are determined by choices made by the modeler, which concern the decay rate and the function mapping from objective to subjective type frequency. The most appropriate choices have not yet been determined in the literature, as it is unclear precisely how long exemplars may persist in memory – particularly for extremely rare words – and precisely how an incoming stream of tokens is filtered to allow only a subset to be presented for potential storage. Until the appropriate choices are elucidated by the literature, we believe it is reasonable to assume that they meet with the caveats outlined in this section, and thus that the overwriting-based and decay-based models are truly (bidirectionally) equivalent.

S5.2.5 Direct comparison to Pierrehumbert (2001)

In Section S5.2.3, we pointed out that the expected behavior (averaged over many runs) of a model in which memory turnover involves exponential decay of exemplars is equivalent to that of one in which memory turnover involves random overwriting of exemplars, provided the former meets the caveats in Section S5.2.4. However, the results of the most well-known exemplar model in the literature (Pierrehumbert, 2001), which uses exponential decay, appear to differ from those of our model, which uses random overwriting. For simulations involving a single phomene category moving under articulatory bias, Pierrehumbert (2001) reported that high-frequency words change faster than low-frequency words, whereas our present model yields no frequency effect. Since Pierrehumbert (2001) has been widely cited as a demonstration that exemplar models necessarily predict frequency effects (that favor high-frequency words), it is important to diagnose the reasons for this difference.

Pierrehumbert (2001) represents the first foray into formal modeling of exemplar dynamics in sound change, and lays important groundwork for our present model. However, as pioneering work, Pierrehumbert's model is necessarily very schematic, and has some limitations. The limitation that is primarily responsible for the discrepancy with the results of our present model concerns the model architecture.

The model presented by Pierrehumbert (2001) does not contain a lexical ("type") level, meaning that – without frequency-based variation in the decay rate or exemplar lifespan – it is technically incapable of obtaining type frequency effects in simulations involving a single phoneme category. Without separate type representations, each phoneme category effectively contains a single type and the same type is necessarily produced on every iteration. Pierrehumbert (2001) observes that the advancement of the type's exemplars is determined by the number of iterations for which the simulation is run: the more iterations, the more the type is produced with articulatory bias, and thus the more it advances. While it is tempting to interpret this observation as reflecting an effect of frequency, it actually reflects an effect of time. This is because each iteration also corresponds to a single application of decay. After sufficient iterations, a specific exemplar will become so weak that its contribution is negligible, meaning it can effectively be dropped from the system. Thus, exemplars have a lifespan, which corresponds to a certain period of time, and each iteration represents a fixed fraction of this lifespan. Both the decay rate and the lifespan of an exemplar are assumed not to vary with type frequency. Therefore, regardless of type frequency, each iteration corresponds to a fixed period of time, and running a simulation for more iterations corresponds to observing a change over a greater period of time. Having the potential to observe an effect of frequency would require the number of iterations taking place in a given period of time to vary with type frequency. This would only be possible in simulations of a single type if decay rate or exemplar lifespan were assumed to vary with type frequency.

In simulations with multiple types within the same category, by contrast, each type may be produced on different numbers of iterations within the same period of time. Our present model,

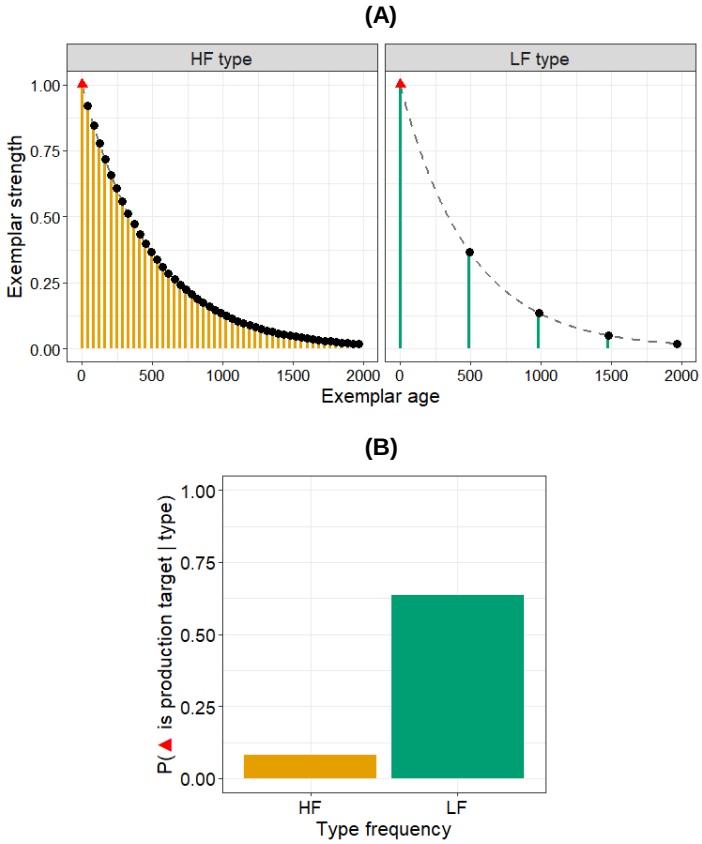

Figure S12: Illustration of how gaps between productions cause high-frequency types (orange; left) to be held back by competition with old exemplars more than low-frequency types (green; right). (A) Comparison of strength of most recent exemplar (red triangle) to strength of older exemplars of the same type (black circles). A high-frequency type has many more old exemplars than a low-frequency type, with correspondingly greater strengths. (B) Probability of selecting the most recent exemplar (red triangle) as the target for production of the type. Since the aggregate strength of old exemplars is greater for a high-frequency type than for a low-frequency type, they compete much more for selection.

which has type-level representations, shows that single-category movement is typically unaffected by type frequency. In a decay-based model that meets the caveats laid out in Section S5.2.4, this lack of frequency effect follows because the strengths of exemplars of a given type continue to decay during the gaps between productions of that type. The shorter the gap, the stronger old exemplars of the type will be relative to the most recent exemplar, and thus the more they will compete with it to provide the acoustic target for the next production of the type. Competition from old exemplars holds a category back, since old exemplars represent earlier (less advanced) stages of the change. Since a high-frequency type has shorter gaps between productions than a low-frequency type, it will be held back by competition from old exemplars more, counterbalancing the fact that it will be produced (with articulatory bias) more often. We illustrate this process in Figure S12, using parameters corresponding to the simulations presented in this paper.

The fact that the model presented by Pierrehumbert (2001) is technically unable to display frequency effects in single-category simulations renders moot the question of differences from our model. While some decay-based models (with type-level representations) *do* display frequency effects in single-category simulations, consistent with the broader suggestions made by Pierrehumbert (2001), these effects are contingent on modeler choices, as discussed in Section S5.2.4. The

literature to date has not recognized this contingency and has taken Pierrehumbert's suggestions extremely generally, giving rise to criticisms that exemplar models necessarily over-predict word frequency effects and cannot explain all the patterns found in empirical studies (Abramowicz, 2007; Dinkin, 2008; Tamminga, 2014; Bermúdez-Otero et al., 2015). As we have shown, these criticisms are not applicable to exemplar models as a class, and the new model presented here is successful in generating all of the reported patterns.

References

Abramowicz, Ł. (2007). Sociolinguistics Meets Exemplar Theory: Frequency and Recency Effects in (ing). *University of Pennsylvania Working Papers in Linguistics*, *13*, 27–37.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as Probability Density Estimation. *Journal of Mathematical Psychology*, *39*, 216–233. doi:10.1006/jmps.1995.1021.

Begg, I. (1974). Estimation of word frequency in continuous and discrete tasks. *Journal of Experimental Psychology*, *102*, 1046–1052. doi:10.1037/h0036356.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, *31*, 297–305. doi:10.3758/BF03194388.

Bermúdez-Otero, R., Baranowski, M., Bailey, G., & Turton, D. (2015). A constant rate effect in Manchester /t/-glottalling: high-frequency words are ahead of, but change at the same rate as, low-frequency words. Paper presented at 2nd Edinburgh Symposium on Historical Phonology.

Blevins, J., & Wedel, A. (2009). Inhibited sound change: An evolutionary approach to lexical competition. *Diachronica*, *26*, 143–183. doi:10.1075/dia.26.2.01ble.

Bowers, J. S. (2000). The modality-specific and -nonspecific components of long-term priming are frequency sensitive. *Memory & Cognition*, *28*, 406–414. doi:10.3758/BF03198556.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707–729. doi:10.1016/j.cognition.2007.04.005.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, 1116. doi:10.3389/fpsyg.2016.01116.

Busemeyer, J. R., Dewey, G. I., & Medin, D. L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 638–648. doi:10.1037/0278-7393.10.4.638.

Carreiras, M., Mechelli, A., & Price, C. J. (2006). Effect of word and syllable frequency on activation during lexical decision and reading aloud. *Human Brain Mapping*, *27*, 963–972. doi:10.1002/hbm.20236.

Chee, M. W. L., Goh, J. O. S., Lim, Y., Graham, S., & Lee, K. (2004). Recognition memory for studied words is determined by cortical activation differences at encoding but not during retrieval. *NeuroImage*, *22*, 1456–1465. doi:10.1016/j.neuroimage.2004.03.046.

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & Psychophysics*, *70*, 604–618. doi:10.3758/PP.70.4.604.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, *108*, 710–718. doi:10.1016/j.cognition.2008.06.003.

Davies, M. (2008-). The Corpus of Contemporary American English: 520 million words, 1990-present. URL: http://corpus.byu.edu/coca/ [Date retrieved: July, 2014].

Diana, R. A., & Reder, L. M. (2006). The low-frequency encoding disadvantage: Word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 805–815. doi:10.1037/0278-7393.32.4.805.

Dinkin, A. J. (2008). The real effect of word frequency on phonetic variation. *University of Pennsylvania Working Papers in Linguistics*, *14*, 97–106.

Ettlinger, M. (2007). An exemplar-based model of chain shifts. In J. Trouvain, & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of the Phonetic Science* (pp. 685–688).

Forster, K. I., & Davis, C. (1984). Repetition Priming and Frequency Attenuation in Lexical Access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 680–698. doi:10.1037/0278-7393.10.4.680.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125. doi:10.1037/0096-1523.6.1.110.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279. doi:10.1037/0033-295X.105.2.251.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming Lexical Neighbors of Spoken Words: Effects of Competition and Inhibition. *Journal of Memory and Language*, *28*, 501–518. doi:10.1016/j.biotechadv.2011.08.021.Secreted.

Harrington, J. (2006). An acoustic analysis of 'happy-tensing' in the Queen's Christmas broadcasts. *Journal of Phonetics*, *34*, 439–457. doi:10.1016/j.wocn.2005.08.001.

Harrington, J., Kleber, F., Reubold, U., Schiel, F., & Stevens, M. (2018). Linking Cognitive and Social Aspects of Sound Change Using Agent-Based Modeling. *Topics in Cognitive Science*, (pp. 1–22). doi:10.1111/tops.12329.

Hintzman, D. L. (1986). "Schema Abstraction" in a Multiple-Trace Memory Model. *Psychological Review*, *93*, 411–428. doi:10.1037/0033-295X.93.4.411.

Hintzman, D. L., & Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology*, *88*, 297–306. doi:10.1037/h0030907.

Hooper, J. B. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie (Ed.), *Current Progress in Historical Linguistics* (pp. 96–105). Amsterdam: North Holland.

Johnson, K., Flemming, E., & Wright, R. (1993). The Hyperspace Effect: Phonetic Targets Are Hyperarticulated. *Language*, *69*, 505–528. doi:10.2307/416697.

Kinoshita, S. (1995). The Word-Frequency Effect In Recognition Memory Versus Repetition Priming. *Memory & Cognition*, *23*, 569–580. doi:10.3758/bf03197259.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*, 262–268. doi:10.3758/BF03193841.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York, NY: Wiley.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238. doi:10.1037/0033-295X.85.3.207.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165–178. doi:10.1037/h0027366.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, *111*, 721–756. doi:10.1037/0033-295X.111.3.721.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*, 357–395. doi:10.1037/0033-295X.115.2.357.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, *47*, 204–238. doi:10.1016/S0010-0285(03)00006-9.

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: a choice model analysis. *Perception & Psychophysics*, *38*, 415–432. doi:10.3758/BF03207172.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–61. doi:10.1037/0096-3445.115.1.39.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 3–27. doi:10.1037/0096-1523.17.1.3.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee, & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam: John Benjamins.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 101–139). Berlin: Mouton de Gruyter.

Pierrehumbert, J. B., & Granell, R. (2018). On hapax legomena and morphological productivity. In *Proceedings of SIGMORPHON 2018*. Stroudsburg, PA: Association for Computational Linguistics.

Schulman, A. I. (1967). Word Length and Rarity in Recognition Memory. *Psychonomic Science*, *9*, 211–212. doi:10.3758/BF03330834.

Shepard, R. N. (1958). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, *65*, 242–256. doi:10.1037/h0043083.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319. doi:10.1016/j.cognition.2013.02.013.

Sóskuthy, M. (2013). *Phonetic biases and systemic effects in the actuation of sound change*. Unpublished doctoral dissertation. University of Edinburgh.

Sóskuthy, M. (2014). Explaining lexical frequency effects: a critique and an alternative account. Paper presented at Sound Change in Interacting Human Systems, 3rd Biennial Workshop on Sound Change, University of California, Berkeley.

Tamminga, M. (2014). Sound Change without Frequency Effects: Ramifications for Phonological Theory. In R. E. Santana-LaBarge (Ed.), *Proceedings of the 31st West Coast Conference on Formal Linguistics* (pp. 457–465). Somerville, MA: Cascadilla Proceedings Project.

Tupper, P. F. (2015). Exemplar Dynamics and Sound Merger in Language. *SIAM Journal on Applied Mathematics*, *75*, 1469–1492. doi:10.1137/140998408.

Wagenmakers, E.-J., Zeelenberg, R., & Raaijmakers, J. G. (2000). Testing the counter model for perceptual identification: effects of repetition priming and word frequency. *Psychonomic Bulletin & Review*, *7*, 662–667. doi:10.3758/BF03213004.

Wedel, A. (2004). Category competition drives contrast maintenance within an exemplar-based production / perception loop. In J. Goldsmith, & R. Wicentowski (Eds.), *Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology* (pp. 1–10). Stroudsburg, PA: Association for Computational Linguistics.

Wedel, A. (2006). Exemplar models, evolution and language change. *The Linguistic Review*, *23*, 247–274. doi:10.1515/TLR.2006.010.

Wedel, A. (2012). Lexical contrast maintenance and the development of sublexical contrast systems. *Language and Cognition*, *4*, 319–355. doi:10.1515/langcog-2012-0018.

Wedel, A., & Fatkullin, I. (2017). Category competition as a driver of category contrast. *Journal of Language Evolution*, (pp. 77–93). doi:10.1093/jole/lzx009.

Wierda, S. M., Taatgen, N. A., van Rijn, H., & Martens, S. (2013). Word Frequency and the Attentional Blink: The Effects of Target Difficulty on Retrieval and Consolidation Processes. *PLoS ONE*, *8*, e73415. doi:10.1371/journal.pone.0073415.

Zipf, G. K. (1935). *The Pscyho-Biology of Language*. Oxford: Houghton Mifflin.