Crosslinguistic transfer as category adjustment: Modeling conceptual color shift in bilingualism

Yevgen Matusevych
Department of Computer Science
University of Toronto
yevgen@cs.toronto.edu

Barend Beekhuizen
Department of Computer Science
University of Toronto
barend@cs.toronto.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
suzanne@cs.toronto.edu

Abstract

We present a general framework for capturing categorical crosslinguistic transfer effects – the influences of linguistic and conceptual categories in a bilingual speaker's languages on each other. By formulating the phenomenon as an instance of cognitive category shift, we achieve a general method for investigating the extent and causes of crosslinguistic transfer in terms of a category similarity space and a set of weighting factors. We apply the model to the well-understood domain of color, formulating transfer as the modulation of conceptual color categories in one language on those of the other language. We analyze the components of the model that predict salient aspects of human data on an observed transfer effect in a range of languages.

Keywords: semantic shift; crosslinguistic transfer; color categories; category adjustment model

Introduction

It is generally agreed that languages in the bilingual mind influence each other (e.g., Jarvis & Pavlenko, 2008; Odlin, 1989). In linguistic terms, there is a bi-directional transfer between the two phonological, semantic, conceptual, etc. systems, and in many cases the two systems tend to converge, or shift towards each other (see examples in Jarvis & Pavlenko, 2008). It is common for second language (L2) learners to apply knowledge of their native language (L1) to the language they learn, but transfer may also occur in the reverse direction. For example, Russian speakers under the influence of their L2 English may stop perceiving the obligatory contrast between *goluboy* ['light blue'] and *siniy* ['dark blue'], a distinction always made by Russian monolinguals (Andrews, 1994).

In domains such as phonology or conceptual semantics, such knowledge is often represented in terms of categories, which may differ across languages. Languages vary widely in the way their words carve up the world into conceptual categories (for an overview, see Malt & Majid, 2013): for example, for the regions of the color space categorized by English speakers as *green* and *blue*, Cantonese uses *luk* ['jade colored'] to refer to parts of *green* and *blue*, and uses *l'ām* ['artificial blue'] for a subregion of English *blue* (Berlin & Kay, 1969). Being bilingual thus typically requires speakers to rely on two – only partially overlapping – sets of categories. For reasons of cognitive efficiency (Kemp et al., 2018), it is natural for the two sets of categories to shift towards each other in the bilingual mind.

Looking at the domain of color terms, for instance, it has been observed that bilinguals' judgments of the best exemplars of a color term in their first language (L1) – the focal members of that color category – are not identical to that of monolingual speakers of the same language. Why this shift in L1 color categories happens is an open question. Presumably, the nature of the L2 conceptual system plays a role in this, but how this second system modulates the first remains unclear.

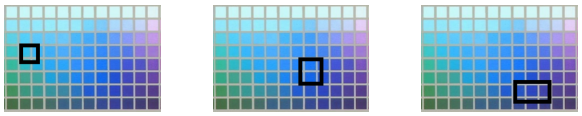
In this paper, we use the well-studied phenomenon of bilingual color shift as a testbed for a new model of categorical crosslinguistic transfer effects. Existing studies show that such effects in the domain of conceptual semantics can be measured and formalized as an instance of category shift due to modulation of the categories in one language by corresponding categories in the other (Ameel et al., 2009; Fang et al., 2016). Here, we develop a novel computational framework that explains such effects in terms of the Category Adjustment Model, which has been used to account for monolingual speakers' behavior in a number of cognitive domains – e.g., spatial (Huttenlocher et al., 1991), phonological (Kuhl, 1991; Feldman, Griffiths, & Morgan, 2009), and color (Bae et al., 2015; Cibelli et al., 2016). Building on this cognitively-natural framework, our new model provides a general method for evaluating the nature and extent of crosslinguistic transfer effects.

Background

Color shift in bilingual speakers

We assume, following standard practice (e.g., Berlin & Kay, 1969), that basic terms in a lexical semantic system reflect a set of conceptual categories of the underlying semantic space – e.g., the term *lán* ['blue'] in Mandarin Chinese refers to a particular category of color (a region of the color space) for Mandarin speakers. However, the same term may encode a somewhat different category in Mandarin–English speakers: that is, conceptual representations of the same color terms can differ in bilingual vs. monolingual speakers. This is an instance of crosslinguistic transfer, in which a bilingual speaker's second language, L2, can influence aspects of their L1, and vice versa. The transfer effects observed in bilingual color terminology, for both L1 and L2 terms, include widening of color categories (i.e., a color term refers to a broader region of perceptual space than in a corresponding monolingual), weakening of contrasts between color categories, increase in the variability of selecting best exemplars for a given color term, etc. (see an overview by Pavlenko et al., 2017).

One effect consistently observed in different bilingual populations (across various L1s and L2s) is conceptual shift: the best exemplars of some color categories appear to be shifted in bilingual speakers compared to monolinguals (Athanasopoulos, 2009; Caskey-Sirmons & Hickerson, 1977, etc.). For example, the Mandarin term *lán* ['blue'] shifts away from



(a) MN (b) MN–EN (c) EN

Figure 1: Focal colors – i.e., best exemplars for given color terms – shown on a color chart (reconstructed from CS&H): (a) *lán* in Mandarin Chinese (MN) monolinguals, (b) *lán* in MN–EN bilinguals, (c) *blue* in English (EN) monolinguals.

the green spectrum (Figure 1a) towards purple (Figure 1c) in bilingual Mandarin–English speakers (Figure 1b).

The shift of L1 color categories was observed in multiple languages by Caskey-Sirmons & Hickerson (1977) (henceforth CS&H), who asked monolingual and bilingual speakers of different languages to name the basic color terms in their L1,[1] and then to select the corresponding focal colors – the best example of each color term – on a color chart. Participants included monolingual speakers of Korean, Japanese, Hindi, Cantonese, and Mandarin Chinese, as well as bilingual speakers of each of those languages as L1, and whose L2 was English. The data for the focal colors of English from monolingual speakers were also gathered in the study. The results suggest that the locations of focal colors in bilinguals' L1 are shifted in color space compared to those in monolinguals of L1. For many color terms this shift occurs towards the focal colors for the corresponding L2 color terms (as identified by monolingual speakers of the L2); see Figure 1. Focal shift has been observed in multiple languages, both in L1 and L2; here we draw on data on L1 shift from CS&H as they investigate the widest range of languages within a single study.

The Category Adjustment Model

Crosslinguistic transfer effects, such as those noted above, can be seen as resulting from an interaction between the conceptual category systems of two languages. We formalize this idea by extending the Category Adjustment Model (CAM; Huttenlocher et al., 1991; Kuhl, 1991). The CAM predicts that humans represent a given perceptual stimulus both as a fine-grained value and as a member of a category. When a speaker is exposed to a stimulus and asked to recollect it, the categorical representation comes into play, shifting the recollected value towards the category prototype. This approach has been used to model effects of linguistic categories on perceptual stimuli – e.g., showing how discrimination of phonetic stimuli is influenced by phonological categories of the language (Feldman et al., 2009).

Recent studies (Bae et al., 2015; Cibelli et al., 2016) show that the CAM can explain some patterns in (monolingual) human color perception as modulated by lexical color terms (which, as noted above, refer to conceptual categories). When participants are shown a color hue (e.g., turquoise) and have to select it on a color wheel, their recollection of the stimulus is shifted towards the focal colors of nearby color categories (in this case, exhibiting influence of both focal green and blue).

Importantly, the above studies show that the recollected representation can be formally predicted from an interaction of the source representations: e.g., the original color stimulus, and the color categories adjacent to that stimulus in the color space. We use this insight as the basis for our formal model of crosslinguistic transfer in bilinguals: a category in one language (e.g., the color category corresponding to an L1 term) is modulated by (nearby) categories in the other language (e.g., the color categories in L2), leading to crosslinguistic influence on the conceptual categories of each language.

Our Model of Bilingual Transfer

Here we present our formal model of crosslinguistic transfer as the modulation of categories in one language of a bilingual by related categories in the other language. We use lexical color terms as the testbed for our model, specifically investigating cases in which bilinguals' L2 color categories influence the perception of their L1 color categories. For reasons of space, we describe our model as it applies to this specific instance of transfer, but the components of the model are generally applicable to any linguistic aspects of language formalizable as categories (e.g., phonological or syntactic categories).

The Bilingual Category Adjustment Model

The example situation we model is as follows: when a bilingual identifies the focal color for a color term in their L1, the selection of that focal color is influenced by the speaker's L2 color categories. Our model operates on two sets of color categories, $f_i \in F$ (the *first* language, L1) and $s_j \in S$ (the *second* language, L2). These color categories correspond to the basic color terms $t_i \in T$ from L1, and $u_j \in U$ from L2. For example, s_j may be a red region of color space denoted by $u_j = red$. These categories are defined in our model using data from monolingual speakers of the L1 and L2. From these, our model makes predictions about how the categories F are shifted in a bilingual speaker of L1–L2: the model predicts p_i, the adjusted category corresponding to f_i, as modulated by the categories S. These predictions are compared to $b_i \in B$, the L1 color categories observed in bilingual speakers of L1–L2.

Each color category, such as f_i, is represented as a three-dimensional normal distribution in the $L^*a^*b^*$ space, a standard representational space that is believed to reflect human perceptual discriminability of colors (Fairchild, 1998). Each such distribution is represented by its mean, e.g., μ_{f_i}, and its variance. We describe below how we estimate the means μ. Because we lack the data to estimate variance for all languages in our study, we make the simplifying assumption that the variances of all color categories are equal (which eliminates this factor from our equations below).

Figure 2 shows how an L1 category, f_i, centered around its mean, μ_{f_i}, can be influenced in a bilingual speaker by the L2 categories, S. Our model predicts the resulting bilingual L1 mean focal color p_i by calculating its mean μ_{p_i} as follows:

[1] Intuitively, basic color terms are monomorphemic lexical items, such as *red* or *purple*, which are psychologically salient to speakers of the language (see Berlin & Kay, 1969, pp. 6–7 for further criteria).

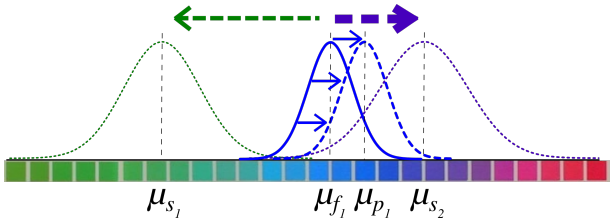


Figure 2: Adjustment of L1 category f_1 by L2 categories s_1 and s_2, with the strength of s_j's influence (arrow thickness) being proportional to the distance between μ_{s_j} and μ_{f_1}. Category p_1 is the predicted category by the model – the shifted f_1. For simplicity, only one dimension of color is shown.

$$E[\mu_{p_i}|\mu_{f_i}] = \mu_{f_i} + \frac{1}{2}\sum_{s_j \in S}[\omega\,(\mu_{s_j} - \mu_{f_i})\,\pi(s_j|f_i)] \qquad (1)$$

Here, we adapt the CAM for multiple categories (Cibelli et al., 2016; Feldman et al., 2009), and reformulate it conceptually to show how L1 category f_i is influenced by the difference between it and each L2 category s_j. We introduce an additional parameter $\omega \in \{-1, 1\}$, which determines the *direction* of influence on μ_{f_i} – i.e., towards μ_{s_j} ($\omega = 1$), or away ($\omega = -1$). The applicability function $\pi(s_j|f_i)$ determines the *degree* of influence, which is greater for categories closer in the color space. We next explain each component of the model.

Defining the color category means (μ)

Color categories, such as f_i, can be defined in either of two ways. First, a category f_i for a color term t_i can be defined based on all the *usages* of t_i – all the colors the speaker has heard referred to by that color term. Second, a category f_i can be based on the *focal* colors – the best exemplars of the term. The centers of the overall color region and of the focal colors are not necessarily the same in people (e.g., Regier, Kay, & Cook, 2005). This means, in our model, we can define μ_{f_i} in terms of usages or focal colors. For the former, the category center μ_{f_i} would be the mean of the sample of all colors labeled by term t_i, and for the latter, it would be the mean of the focal colors identified by people for term t_i.

Here, we assume the L1 categories to be centered around their focal colors, because the data we are modeling consist of the naming of focal colors – that is, we want to model how L1 focal colors shift in a bilingual (b_i) compared to a monolingual (f_i). We thus set μ_{f_i} and μ_{b_i} based on focal colors, estimating these values using the focal color naming data in CS&H.

For our modeling of L2, the situation is more complex. Because focal colors are not necessarily the center of a color region, bilinguals have to learn two properties of an L2 color term: both the region of color covered by the term, *and* what constitutes the focal color of that term. We do not know in our modeling which of these L2 category representations influences a bilingual's recollection of their L1 categories. Thus we define μ_{s_j} to be a weighted average of the two:

$$\mu_{s_j} = \alpha\,\upsilon_{s_j} + (1-\alpha)\phi_{s_j} \qquad (2)$$

where υ_{s_j} is the center of the color region referred to by all usages of the L2 term u_j; ϕ_{s_j} is the center of the focal colors of u_j; and α is a parameter fitted to the data as described below.

Weighing the category influence (π)

The function $\pi(s_j|f_i)$ in Eqn. (1) indicates the degree to which category s_j influences f_i in predicting the adjusted category p_i. Intuitively, more similar L2 categories are expected to influence the L1 category more than less similar ones. We take the influence to decay exponentially with the increase of Euclidean distance d between category means μ_{f_i} and μ_{s_j}:[2]

$$\pi(s_j|f_i) = \exp(-c\,d(\mu_{f_i}, \mu_{s_j})) \qquad (3)$$

where c is a constant fitted to the data, known as the sensitivity parameter. Intuitively, c determines how sharply the degree of influence of an s_j drops off as it is further from f_i.

Translation equivalence (c and ω)

The influence of s_j on f_i may also be affected by their corresponding lexical terms, u_j and t_i. Specifically, u_j may be a *translation equivalent*[3] of the L1 term t_i ($t_i \widehat{=} u_j$; e.g., Mandarin *lán* and English *blue*). Following proposals on transfer effects through translation equivalents (e.g., Degani, Prior, & Tokowicz, 2011), we propose that such an s_j may influence the target L1 category differently than other L2 categories – quantitatively and/or qualitatively.

First, the difference may be one of *degree*: the L2 category of a translation equivalent may influence the L1 category more than do other L2 categories. We implement this idea by letting c in the applicability function π (Eqn. 3) take on two different values: c_{TE} (for the translation equivalent) and $c_{\neg\text{TE}}$ (for other L2 categories). Because c_{TE} and $c_{\neg\text{TE}}$ are fitted to the data independently, the model can establish a greater degree of influence for translation equivalents.

Second, the difference may be one of the *kind* of influence: while category adjustment models have been limited to *attraction* between categories, crosslinguistic transfer can also lead to *repulsion* effects (e.g., Athanasopoulos, 2009). Parameter ω in Eqn. (1) implements the idea of attraction and repulsion, by moving f_i towards ($\omega = 1$) or away from ($\omega = -1$) *all* categories s_j. Now, we go a step further and let the value of ω be set differently for translation equivalents and non-translation equivalents. In particular, we see whether the model can achieve a better fit to human data if it can have the category s_j of the translation equivalent for f_i attract f_i ($\omega_{\text{TE}} = 1$), while the other categories in S repel f_i ($\omega_{\neg\text{TE}} = -1$).

Set-up for Computational Experiments

We first describe how we estimate the means of the color categories that serve as input to our model. Then we explain the setting of parameters, which gives rise to a number of model variants that we investigate.

[2] $\pi(s_j|f_i)$ values for f_i over $s_j \in S$ are normalized to sum to 1.

[3] The term commonly used in literature on bilingualism; in our case the meanings of color terms are not strictly equivalent. Here, we determine translation equivalents by the glosses in CS&H.

Estimating color categories (μ) from human data

The L1 data we model are focal colors for basic color terms. CS&H collected such data from all the L1 and L2 languages in their study, reporting these in charts of the type shown in Figure 1.[4] We use these charts to reconstruct the $L^*a^*b^*$ coordinates of the average focal color selected for each color term by the various populations of speakers. We estimate μ_{f_i} input to our model for each L1 using the focal colors identified by monolinguals in these languages (Korean, Japanese, Hindi, Cantonese, and Mandarin Chinese). We estimate μ_{b_i} for each L1 (the human data we match our model predictions against) using the focal colors identified by bilingual Korean–English, Japanese–English, etc. speakers.

Recall that for the L2 (English) color categories S, we calculate each μ_{s_j} in our model as a mixture of two influences: ϕ_{s_j} is the center of the focal colors for the L2 term u_j (analogous to μ_{f_i}, μ_{b_i} above), while υ_{s_j} is the center of the full color region referred to by u_j (see Eqn. 2). For the focal color estimate, ϕ_{s_j}, we use the CS&H data of monolingual English speakers, and augment this with similar English focal color data available from Berlin & Kay (1969) and Sturges & Whitfield (1995) to increase accuracy. The means of the full English color category regions, υ_{s_j}, are obtained from publicly available color naming data[5].

Parameter setting and model variants

We investigate several model variants which are determined by the following parameters:

- $\alpha \in \{0.0, 0.1, ...1.0\}$: determines whether L2 colors are defined by all usages and/or focal colors (Eqn. 2).
- $c \in \{0.01, 0.02, ..., 1.00\}$: specifies the relative degree of bias of L2 colors on L1 colors (Eqn. 3).
- $\omega \in \{-1, 1\}$: determines the direction of the bias (attraction or repulsion) of L2 on L1 (Eqn. 1).

In each model variant, we fix some parameters, and set any others in a leave-one-out procedure – specifically, by optimizing them jointly using a grid search on four languages and using the best values in the simulation of the fifth.

Model variants FOC, USG, and MIX. We start with basic models which have no effect of translation equivalence; i.e., we find a single optimal value for c, and set $\omega = 1$ (all categories attract f_i). When $\alpha = 0$, we have a model variant FOC in which the mean of the focal colors defines each μ_{s_j}. Conversely, when $\alpha = 1$, we have a model variant USG in which the mean of all usages defines μ_{s_j}. Finally, we have model MIX in which we find the optimal α to weigh these.

Model variants FOC-TE$_{\text{DEG}}$, USG-TE$_{\text{DEG}}$, and MIX-TE$_{\text{DEG}}$. We build on the basic models above by letting translation equivalents vs. other terms have a different *degree* of influence. Instead of a single c value, we fit c_{TE} and $c_{\neg\text{TE}}$ to the data independently.

[4]The number of color terms varied between 4 and 8 per language; *white*, *gray*, and *black* were excluded from analysis in CS&H.

[5]https://blog.xkcd.com/2010/05/03/color-survey-results/

Model variants FOC-TE$_{\text{DIR}}$, USG-TE$_{\text{DIR}}$, and MIX-TE$_{\text{DIR}}$. We build on the basic models by letting translation equivalents vs. other terms have a different *direction* of influence. Instead of a single $\omega = 1$ value, we set $\omega_{\text{TE}} = 1$ and $\omega_{\neg\text{TE}} = -1$.

Model variants FOC-TE$_{\text{BOTH}}$, USG-TE$_{\text{BOTH}}$, and MIX-TE$_{\text{BOTH}}$. These translation equivalence variants combine the effects of the DEG and DIR variants.

Results

The input to the model consists of color terms in L1 and L2 ($t_i \in T$ and $u_j \in U$, respectively) together with the corresponding category means (μ_{f_i} and μ_{s_j}) as defined above using monolingual data. For each L1 color term t_i, the model outputs the predicted μ_{p_i} of the corresponding focal color in an L1–L2 bilingual. Each prediction p_i of the model is compared to $b_i \in B$, the set of observed bilingual color categories for L1 – the human data our model is evaluated against.

Evaluation of computational experiments

We measure the error in the fit of the model to the human data as the average Euclidean distance d between the focal color μ_{p_i} predicted by the model and the focal color μ_{b_i} identified by bilinguals in CS&H (the lower the d, the better the model fit). We compare each of the model variants to a baseline, BASE, that predicts the focal color of each bilingual L1 category μ_{p_i} to be equivalent to μ_{f_i}. This baseline assumes that a bilingual's L1 categories are equivalent to those of a monolingual – i.e., there is no shift arising from an influence of L2.

Table 1 presents the performance of each model variant, in average distance d of its predictions across all terms and languages. The table also reports the percentage improvement over the baseline, and the β coefficient of a mixed-effects regression fitted to the data, which shows the degree to which each model is better than the baseline.[6] All model variants show a significant improvement of 18–24% over the baseline.

Comparison of model variants

First, consider the model variants based on whether the definition of the L2 categories is given by focal color (FOC), all usages (USG), or an optimal mixture (MIX). We find that the the model can achieve a match to human behavior that outperforms the baseline by 22% with the USG variant, which has a single optimized parameter (c). The other variant that outperforms the baseline, MIX, uses an additional optimized parameter (α), but its improvement over USG is not significant. The model variant FOC, and all its translation equivalence variants, do not outperform the baseline. This finding suggests that, for the bilingual speakers in the simulated population, the L1 focal colors are influenced by full L2 color regions (as in the USG model) rather than by L2 focal colors.

[6]Specifically, this regression introduces 12 binary dummy predictors (one per model, with BASE being a reference level) and fits a number of parallel hyperplanes (one per color term) to the d values of all models, this way testing the difference between each model and BASE, while taking into account the variation among L1 color terms.

Table 1: Model error as d averaged over all color terms (distance of prediction to human data).

	d	$\Delta E, \%^{\dagger}$	$\beta^{\ddagger}$
BASE	19.4	—	19.4
FOC	15.5	19.8	-3.8^{*}
USG	15.0	22.3	-4.3^{**}
MIX	14.7	24.2	-4.6^{***}
FOC-TE_{DEG}	15.6	19.7	-3.8^{*}
USG-TE_{DEG}	15.0	22.4	-4.3^{***}
MIX-TE_{DEG}	14.8	23.4	-4.5^{***}
FOC-TE_{DIR}	15.7	19.2	-3.7^{*}
USG-TE_{DIR}	15.1	22.0	-4.2^{**}
MIX-TE_{DIR}	14.7	24.2	-4.6^{***}
FOC-TE_{BOTH}	15.8	18.7	-3.6^{*}
USG-TE_{BOTH}	14.9	22.9	-4.4^{**}
MIX-TE_{BOTH}	14.8	23.5	-4.5^{***}

† ΔE is the percentage improvement in error rate over the baseline.
‡ β is the standardized regression coefficient in the mixed-effects regression fitted to the error terms d (per color) in all 13 models.
$^{*}, ^{**}, ^{***}$ Significantly better than BASE at .05, .01, and .001 level, respectively; all p-values are Bonferroni-corrected.

Second, consider the more complex model variants that encode crosslinguistic transfer effects (-TE_{DEG}, -TE_{DIR}, -TE_{BOTH}). Somewhat surprisingly, none of these significantly outperform the simpler models. On the surface, this suggests that translation equivalents may not have a special status during color shift. However, while only the parameter c is optimized in our simple model USG, on inspection we see that the optimal value of c ensures that each L1 category is substantially affected solely by its L2 nearest neighbor. Moreover, it turns out that, for all L1 terms in all languages in this study, the L2 nearest neighbor *is* the translation equivalent. Thus, there is no need to tune separate parameter values for translation equivalents vs. others: the simpler model already captures the special status of translation equivalents with the single parameter c. This finding suggests that (again, for this population) the influence of L2 on L1 focal colors is primarily limited to the L2 color categories of the translation equivalents (which are the L1 colors' nearest neighbors).

In combining our two findings, a picture thus emerges of an L1 focal color being nudged in the direction of the center of a color region covered by all usages of its translation equivalent in L2, rather than by the corresponding L2 focal color. Further research is needed to flesh out this picture considering other languages and other populations of bilinguals. First, we note that it is possible that the bilingual speakers in this study had simply not sufficiently learned the locations of the focal colors of their L2 for those to influence their L1. Second, a close L2 category that is not a translation equivalent of L1 may have an influence, but this sample of languages and terms confounds those two properties.

Finally, while our model is a general framework that can incorporate a variety of influences on category shift, in this work we have only considered the factors of color category

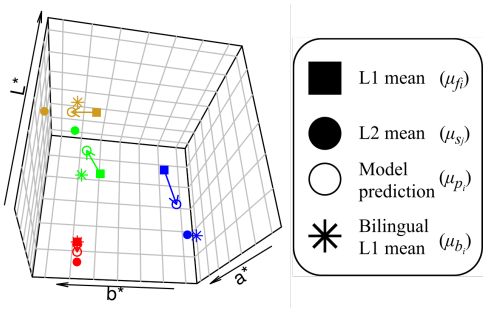


Figure 3: Human data and model predictions for Mandarin color terms. For *yellow*, μ_{p_i} is close to μ_{b_i}. For *blue*, the direction is correct, but the model 'undershoots' μ_{b_i}. For *green*, μ_{b_i} is equally far from μ_{f_i} and μ_{p_i}. For *red*, $\mu_{b_i} = \mu_{f_i}$, while μ_{p_i} is further away.

similarity and term translation. Clearly these are not the only potential influences on crosslinguistic transfer. For example, Figure 3 illustrates the human data and the predictions from our model for Mandarin color terms. This figure shows that some L1 colors do not shift in bilinguals (here, *red*), and that some colors shift in a somewhat different direction than toward the L2 translation equivalent (here, *green*). Future work will also need to consider the conceptual and linguistic properties that bring about such patterns.

Discussion

We present a framework for formalizing crosslinguistic transfer effects in bilingualism as a domain-general mechanism of category shift. We extend the Category Adjustment Model (Huttenlocher et al., 1991; Feldman et al., 2009; Cibelli et al., 2016) to apply to conceptual and linguistic categories *across* languages. This achieves a general method for investigating the extent and causes of bilingual transfer in terms of a category similarity space and a set of weighting factors. To test our model, we focus on the well-understood domain of color. Color terms across languages are associated with varying conceptual categories of color (e.g., Berlin & Kay, 1969), and the color categories of an L1 can influence those of the L2, and vice versa, in bilingual speakers (see Pavlenko et al., 2017). Here we model the influence of L2 on L1.

Our model outperforms the baseline in matching bilingual human data on color naming by 22–24%, shedding light on the factors that influence crosslinguistic transfer in this domain. The model variant in which an L1 color category is biased towards the center of the full L2 color region performs better than the one in which the bias is directed towards the corresponding L2 focal color. This suggests that learners of L2 conceptual categories rely on the full range of language usage events which map the name of the target category to a real-world referent – at least before they have a clear intuition about the category prototype. The best model variant also gives primary influence to the L2 category which is closest to the L1 target, with other L2 categories having negligible effect. Because the closest L2 category was always the translation of the L1 term, we could not consider translation equivalence

of terms as an independent factor. In addition to further exploration of this factor, we need to explore other influences in our framework of linguistic or conceptual properties (such as phonological similarity or word frequency) that may modulate the positioning of conceptual categories in bilinguals.

Interestingly, letting non-translation equivalent L2 categories *repel* the L1 category did not improve model performance. This contrasts with the effect that Greek–English speakers shift their representation of *ble* ['blue'] towards English *blue*, but their representation of *ghalazio* ['light blue'] shifts in the opposite direction (Athanasopoulos, 2009). While our results suggest that such effects do not apply to all color categories, the inability to capture such influences points to a limitation of our model: each L1 category is independently affected by L2 categories. But in reality (as in the Greek case), L1 categories do not shift in isolation: color systems influence each other as a whole.

Embedding our model in a learning framework may help to address such "system-wide" effects. At present, our model only considers two time points – the beginning and end of L2 learning – as is common when comparing L2 learners to a monolingual reference group (e.g., Schmid & Dusseldorp, 2010). But system-wide changes may arise from incremental adjustments in which shifts in one category bring about changes that lead to shifts in another. An important question is to consider how a learning model could affect conceptual shift over time, as this would allow for modeling the underlying sources of conceptual shift. Our preliminary simulations show that one good candidate is a statistical learning model based on the mixture of Gaussians (e.g., McMurray, Aslin, & Toscano, 2009). Operationalizing the shift as a gradual process would also resolve a theoretical issue: our current model assumes some "end state" of L2 acquisition, whereas bilingual acquisition typically is ongoing (Larsen-Freeman, 2006). In this respect, taking into account speakers' L2 proficiency (not reported by CS&H) would be another important step forward.

Acknowledgments: Supported by NSERC RGPIN-2017-06506.

References

Ameel, E., Malt, B. C., Storms, G., & Van Assche, F. (2009). Semantic convergence in the bilingual lexicon. *Journal of Memory and Language*, *60*, 270–290.

Andrews, D. R. (1994). The Russian color categories *sinij* and *goluboj*: An experimental analysis of their interpretation in the standard and émigré languages. *Journal of Slavic Linguistics*, *2*, 9–28.

Athanasopoulos, P. (2009). Cognitive representation of colour in bilinguals: The case of Greek blues. *Bilingualism: Language and Cognition*, *12*, 83–95.

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *JEP: General*, *144*, 744–763.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. UC Press.

Caskey-Sirmons, L. A., & Hickerson, N. P. (1977). Semantic shift and bilingualism: Variation in the color terms of five languages. *Anthropological Linguistics*, *19*, 358–367.

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The Sapir-Whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PLOS ONE*, *11*(8), e0161521.

Degani, T., Prior, A., & Tokowicz, N. (2011). Bidirectional transfer: The effect of sharing a translation. *Journal of Cognitive Psychology*, *23*, 18–28.

Fairchild, M. D. (1998). *Color appearance models*. Addison-Wesley.

Fang, S.-Y., Zinszer, B. D., Malt, B. C., & Li, P. (2016). Bilingual object naming: A connectionist model. *Frontiers in Psychology*, *7*, 644.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*, 752–782.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, *98*, 352–376.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Attention, Perception, & Psychophysics*, *50*, 93–107.

Larsen-Freeman, D. (2006). Second language acquisition and the issue of fossilization: There is no end, and there is no state. In Z. Han & T. Odlin (Eds.), *Studies of fossilization in second language acquisition.* Multilingual Matters.

Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *WIREs Cognitive Science*, *4*, 583–597.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Sci.*, *12*, 369–378.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press.

Pavlenko, A., Jarvis, S., Melnyk, S., & Sorokina, A. (2017). Communicative relevance: Color references in bilingual and trilingual speakers. *Bilingualism: Language and Cognition*, *20*, 853–866.

Regier, T., Kay, P., & Cook, R. S. (2005). Universal foci and varying boundaries in linguistic color categories. In *Proc. CogSci.*

Schmid, M. S., & Dusseldorp, E. (2010). Quantitative analyses in a multivariate study of language attrition: the impact of extralinguistic factors. *Second Lang. Res.*, *26*, 125–160.

Sturges, J., & Whitfield, T. (1995). Locating basic colours in the Munsell space. *Color Research and Application*, *20*, 364–376.