Bilinguals exhibit semantic convergence while maintaining near-optimal efficiency

Maya Taliaferro[*,1] **(mjt10029@nyu.edu)** **Nathaniel Imel**[*,1] **(n.imel@nyu.edu)**

Esti Blanco-Elorrieta[§,1,2] **(eb134@nyu.edu)** **Noga Zaslavsky**[§,1] **(nogaz@nyu.edu)**

[1] Department of Psychology, New York University
[2] Department of Neural Science, New York University
[*] Equal lead contribution; [§] Equal senior contribution

Abstract

Systems of semantic categories vary across languages, but this variation appears to be constrained by pressure for optimizing a complexity-accuracy tradeoff known as the Information Bottleneck (IB) principle. This finding, however, has been based primarily on individual languages and it remains largely unknown how bilinguals navigate the category systems of two different languages, particularly when these languages' category boundaries do not overlap. Here, we address this gap in the literature by combining theory-driven experiments with an extension of the IB framework to bilinguals. Specifically, we investigate bilingual vs. monolingual category boundaries in English and Mandarin via a two-alternative forced-choice (2AFC) labeling task on six continua that interpolate between two distinct everyday objects (e.g., plate and bowl). We find that: (1) bilinguals do not maintain two monolingual-like systems but rather exhibit a converged semantic system influenced equally by both languages; and (2) this departure from monolinguals is nonetheless constrained by the same pressure for efficiency that operates in monolinguals. These findings provide new insight into how bilinguals navigate cross-linguistic semantic variation and suggest that despite having to accommodate myriad sociolinguistic factors, a drive for efficiency is also a key factor that shapes bilingual category systems.

Keywords: bilingualism; communicative efficiency; conceptual category systems; information theory

Introduction

The world is filled with a vast array of entities, from tangible objects like chairs and trees to intangible concepts like emotions and ideas. To make sense of this complexity and communicate effectively about it, humans classify these entities into distinct categories based on shared features (Rosch & Mervis, 1975; Medin & Schaffer, 1978) that they then convey to others via words. Importantly, humans do not assign unique labels to every possible variation of perceivable features; instead they encode only a limited subset (for review, see Malt & Majid, 2013; Malt, 2024). Notably, different cultures and languages prioritize different features to establish their category boundaries, resulting in segmentations that can differ significantly across languages (Berlin & Kay, 1969; Van Hell & De Groot, 1998; Malt, 1995). For example, consider the representation of "chair" and "sofa" for Mandarin versus English speakers. In English, both a cushioned seat for one person and a wooden seat share the same name ("chair"). However, Mandarin speakers group the cushioned seat under the same name as a multi-person cushioned seat—what English speakers would refer to as a "sofa" (Malt, Sloman, & Gennari, 2003). In other words, English uses the number of people that can be seated as the defining feature, but Mandarin uses the material of the seat instead.

It has long been held that these features that define the boundaries of word meanings are not random, but based on the communicative constraints imposed by evolving culture-specific needs, language contact, and the morphological resources of the language (Malt, 1995; White, Storms, Malt, & Verheyen, 2018). More recently, it has been argued that cross-linguistic semantic variation is constrained by a need to support efficient communication (Kemp, Xu, & Regier, 2018; Zaslavsky, Kemp, Regier, & Tishby, 2018). In particular, Zaslavsky et al. (2018) argued that semantic systems evolve under pressure to efficiently compress meanings into forms via a general information-theoretic optimality principle known as the Information Bottleneck (IB; Tishby, Pereira, & Bialek, 1999), providing a theoretical framework which has been supported by broad empirical evidence across languages and semantic domains (Zaslavsky et al., 2018; Zaslavsky, Regier, Tishby, & Kemp, 2019; Zaslavsky, Maldonado, & Culbertson, 2021; Mollica et al., 2021). This suggests that, in combination with sociolinguistic and possibly other cultural factors, a drive for efficiency shapes the ways languages name (or label) objects.

However, this body of literature has mostly focused on monolingual speakers, leaving open two important questions related to bilinguals, who in order to communicate must navigate not only multiple grammars and phonetic inventories, but also differing systems of sematic categories.

The first open question is how bilinguals integrate category knowledge from more than one language. Theoretically, they could address this challenge using one of two broad strategies. First, bilinguals could maximize alignment to the naming conventions of each language by developing two separate "monolingual-like" systems to alternate between. Alternatively, bilinguals might conserve resources by converging to a shared system (see Ameel, Storms, Malt, & Sloman, 2005; Ameel, Malt, Storms, & Van Assche, 2009; White, Malt, & Storms, 2017), in accordance with a common mechanism account of bilingualism (Blanco-Elorrieta & Caramazza, 2021). These two possibilities make distinct predictions about bilinguals' category boundaries. Under the first hypothesis, bilinguals' boundaries would align with the monolingual's boundaries in each language, even when the two monolingual systems have different boundaries. Under the second hypothesis,

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

bilinguals' boundaries in both languages would converge, i.e., tend toward each other and deviate from the two monolingual systems. This convergence could take different shapes based on the influence of each language. If both languages exert equal influence, one could predict that this converged boundary will fall midway between the monolingual boundaries on either end. If one language dominates, the prediction would be that bilinguals' boundaries will shift towards one another, but will be biased toward the dominant language.

The second open question is to what extent the way bilinguals integrate category knowledge is shaped by pressure for efficiency. On the one hand, if pressure for efficiency is indeed a fundamental cognitive constraint, as implied by prior work (Kemp et al., 2018; Gibson et al., 2019; Zaslavsky, 2020), then we should expect that bilinguals would be as efficient as monolinguals, regardless of their need to accommodate additional (socio)linguistic constraints. On the other hand, such sociolinguistic constraints may compete with and outweigh the need for efficiency and potentially pull the bilingual category systems in directions that are less efficient.

Prior work on container naming in Dutch and French suggests that bilingual category systems may converge (Ameel et al., 2005, 2009; White et al., 2017). There is also preliminary evidence that both bilinguals and monolinguals in this domain may be near-optimally efficient in the IB-sense (Zaslavsky et al., 2019, and see also Xu, Regier, & Malt, 2016 for a different approach to studying efficiency in this domain). However, the container stimuli at the basis of these studies only sparsely cover the full semantic domain and therefore do not allow teasing apart precise category boundaries and fully testing the theoretical predictions of the IB framework.

Here, we take a different approach to studying the bilingual lexicon and focus on two typologically distinct languages, English and Mandarin. We present a two-alternative forced-choice (2AFC) labeling task on six continua that interpolate between two everyday objects, as shown in Figure 2. These stimuli were carefully designed to control for variation in the monolingual boundaries, capturing both cases in which the boundaries overlap in English and Mandarin and cases in which they differ. Our results reveal that (1) bilinguals do not maintain distinct monolingual-like systems; instead, they exhibit a converged category system equally shaped by both languages; and (2) the bilinguals' departure from monolinguals is nonetheless constrained by the same pressure for efficiency that operates in monolinguals. These findings provide new insight into how bilinguals navigate cross-linguistic semantic variation and suggest that pressure for efficiency also operates in bilinguals, distinctly from sociolinguistic factors.

How do bilinguals integrate two different category systems?

We begin here with presenting our experimental paradigm and empirical data for addressing our first question, and in the following section we address the second question by conducting an efficiency analysis with respect to these data.

To identify the categorical boundaries in the naming patterns of everyday artifacts and determine how they differ between monolinguals and bilinguals, we conducted an online 2AFC task (Figure 1). Participants in the task provided informed consent and were compensated for their time. The study was approved by the institutional review board.

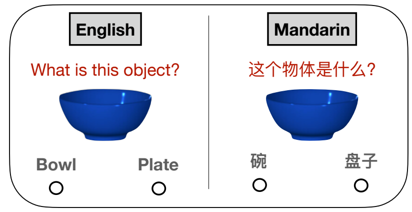

Figure 1: Example of the 2AFC task as seen by participants.

Stimuli

We chose pairs of related everyday objects (e.g., a plate and a bowl) and created a continuum from one object to the other by manipulating their visual features (e.g., depth) in a gradual, step-wise manner. We created a total of 15 continua, each made up of 7 steps which were made as realistic, 500x500px 3D images using Blender.

Subsequently, we normed these continua to obtain the most frequent category names for each item in both languages. The norming procedure consisted of 30 monolingual English (19 females, 11 males; $M_{\text{age}} = 37.3, SD = 11.0$) and 30 monolingual Mandarin (16 females, 14 males; $M_{\text{age}} = 30.9, SD = 6.7$) participants naming the object displayed on the screen. We used the first (step 1), middle (step 4) and last (step 7) item of each continuum for this task. The most frequent names for the first and last items (e.g., "plate" and "bowl") determined the categories of each continuum. We used the middle (step 4) item naming procedure for two purposes. First, it allowed us to reject any continuum where there was a third category between our two ends (e.g., in continuum from 'couch' to 'chair', a blended image in the middle would be called a 'loveseat'). Second, the name distribution for the middle object allowed us to assess whether the category boundary change (e.g., when a "bowl" beings to be perceived as a "plate") occurred at the same step or at a different step across languages. Based on this, we chose 3 continua where the category boundaries overlapped for both languages (referred to as 'same continua'), and another 3 where the category change occurred at a different step (referred to as 'different continua'). Figure 2 shows the final 'different' (a) and 'same' (b) continua.

Online 2AFC task

Participants. We recruited 48 monolingual English speakers (32 females, 14 males, 2 non-binary/prefer not to say; age: $M_{\text{age}} = 40.7, SD = 14.6$; $M_{\text{proficiency}} = 6.67$; $M_{\text{AoA}} = 0.14$ years), 46 monolingual Mandarin speakers (24 females, 21 males, 1 non-binary/prefer not

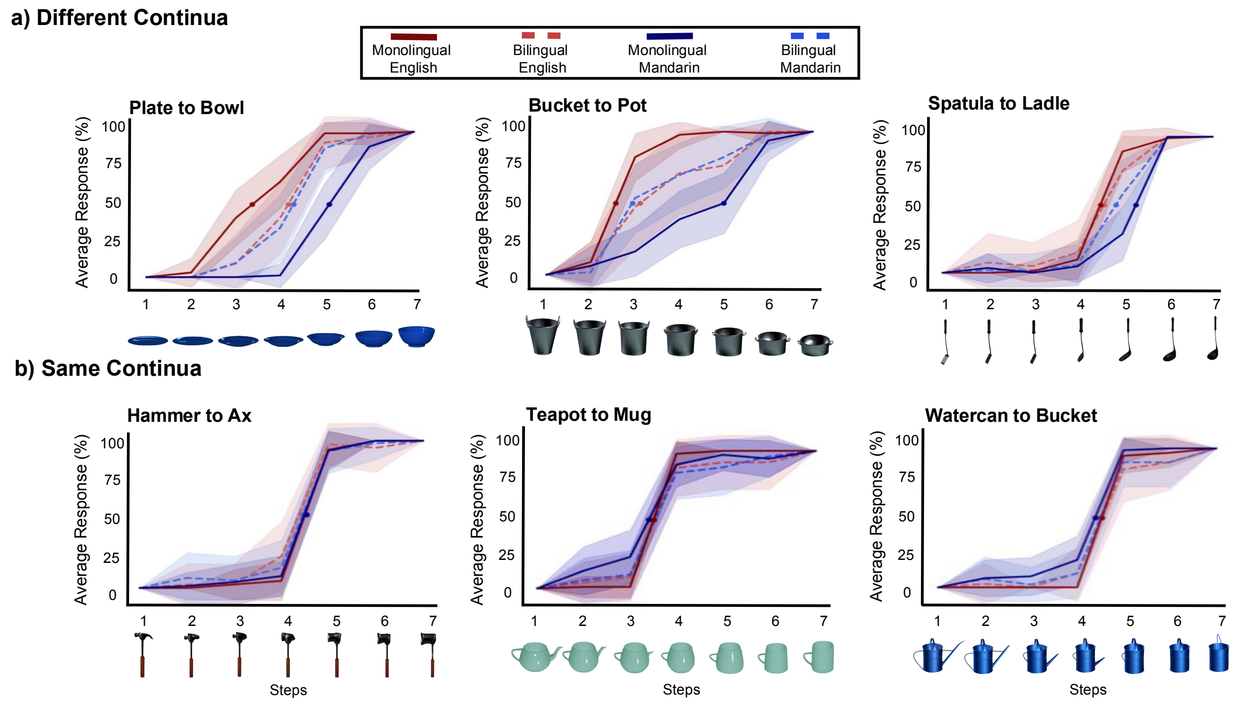

Figure 2: Behavioral results. Curves show the averaged labeling patterns for each language group for (a) 'different boundary' and (b) 'same boundary' continua. The x-axis represents the item step number, and the y-axis shows the proportion of second category responses (e.g., "Bowl" in the Plate-to-Bowl continuum). Shaded regions indicate 95% confidence intervals.

to say; $M_{\text{age}} = 32.3, SD = 8.9$; $M_{\text{proficiency}} = 6.12$; $M_{\text{AoA}} = 0.47$ years) and 34 Bilingual speakers (20 females, 14 males; $M_{\text{age}} = 21.3, SD = 5.4$; $M_{\text{proficiencyMandarin}} = 6.09, M_{\text{proficiencyEnglish}} 5.75$; $M_{\text{AoAMandarin}} = 0.64$, $M_{\text{AoAEnglish}} = 6.21$) for this study. Six of these bilinguals reported being simultaneous learners of English and Mandarin while the rest were Mandarin first sequential learners ($M_{\text{AoAMandarin}} = 0.03$, $M_{\text{AoAEnglish}} = 7.55$). All participants were recruited online and lived either in the USA (monolingual English and bilinguals) or China (monolingual Mandarin). Participants rated their own ability in listening, speaking, reading, and writing on a scale from 1 ('Very poor') to 7 ('Excellent'). We used the average of these four ratings as their proficiency score. Monolinguals who reported significant proficiency in a second language (above 3 out of 7) or bilinguals who were proficient in a third language and/or not highly proficient in both English and Mandarin were excluded from our dataset.

Procedure. Participants saw one image with the two labels corresponding to both ends of the continua (e.g., "Plate" and "Bowl" from Plate-to-Bowl continua) and selected the label that best matched. (Figure 1). Each object was presented twice in random order. All instructions and communication occurred entirely in the language of the experiment. Bilingual participants completed the task in both English and Mandarin, with sessions at least 24 hours apart. We counterbalanced the test order across bilinguals.

Behavioral Analysis. We defined the categorical boundary as the interpolated step in each continuum where participants were equally likely to choose either category label. Using linear interpolation, we estimated this point for each language group and compared group differences with Welch's two-sample t-tests for: (a) monolingual English vs. monolingual Mandarin, (b) bilingual English vs. monolingual English, (c) bilingual Mandarin vs. monolingual Mandarin, and (d) bilingual English vs. bilingual Mandarin. We excluded participants who failed to consistently label the unambiguous exemplars (steps 1 and 7).

Convergence of bilingual categorization

Our behavioral analysis tested competing theories of bilingual categorization – whether bilinguals maintain separate monolingual-like semantic systems or form a shared, blended conceptual system across both languages. Our results strongly support a shared system: while English monolinguals and Mandarin monolinguals showed a significant difference in crossover step in the 'different continua' ($t = -9.02$, $p < 0.001$, CI $= [0.15 : 0.24]$), bilinguals' crossover steps were not different in Mandarin and English ('different continua': $t = 0.46, p = 0.64$, CI $= [-0.04 : 0.07]$). Importantly, neither the monolinguals nor the bilinguals showed a difference in the control, 'same continua' (Monolingual English vs Mandarin: $t = 0.04$, $p = 0.97$, CI $= [-0.06 : 0.06]$; bilinguals in English vs Mandarin: $t = -1.13$, $p = 0.25$, CI $= [-0.07 : 0.02]$). Thus, these findings confirm that the categorization representations for ‘different continua’ differ across languages but that bilinguals rely on a unified representation across both of their languages. Furthermore, we found that both languages exert equal influence on this converged boundary in our bilinguals, as their category boundaries were significantly different from the monolingual boundaries of each language (monolingual English vs. bilingual English: $t = -2.74$, $p < 0.01$, CI $= [-0.13 : -0.02]$; monolingual Mandarin vs. bilingual Mandarin: $t = 4.42$, $p < 0.001$, CI $= [0.11 : 0.36]$). An exploratory analysis excluding the 6 simultaneous bilinguals still revealed a significant difference from monolingual boundaries. Taken together, these findings provide evidence that bilinguals utilize a converged categorization system influenced equally by both languages (for a breakdown of ‘Different continua’ results, see Table 1).

Continua	MonoEng vs. MonoMan	MonoEng vs. BilingEng	MonoMan vs. BilingMan	BilingEng vs. BilingMan
Plate to Bowl	$p < 0.001$***	$p < 0.05$*	$p < 0.01$**	$p = 0.81$
Bucket to Pot	$p < 0.001$***	$p < 0.01$**	$p < 0.01$**	$p = 0.88$
Spatula to Ladle	$p < 0.05$*	$p = 0.91$	$p = 0.32$	$p = 0.34$

Table 1: p values from Welch’s t-test on ‘different continua’ for the pairwise comparisons between language groups. None of the comparisons for the ‘same continua’ were significant.

Are bilingual category systems efficient?

The empirical results above show that bilinguals can differ from monolinguals in both languages with regard to their categorization patterns. But to what extent does this departure from the monolinguals support efficient communication? While prior work has shown that monolingual category systems (divergent or not) are typically highly efficient, it remains unknown whether the bilingual patterns are equally constrained by efficiency. For example, if the semantic convergence in the bilingual patterns is driven by sociolinguistic pressures, then it may lead bilinguals to compromise efficiency while satisfying other constraints. On the other hand, if pressure for efficiency is a fundamental cognitive bias, then this bias may constrain the bilingual category systems in a similar manner as it constrains monolinguals.

Theoretical framework

To assess the efficiency of the converged bilingual category systems, we turn to the theoretical framework of Zaslavsky et al. (2018), which argues that speakers efficiently compress meanings into labels according to a general information-theoretic principle called the Information Bottleneck (IB; Tishby et al., 1999). This framework is particularly relevant in our context because its characterization of semantic systems across languages has gained broad empirical support (Zaslavsky et al., 2018, 2019; Mollica et al., 2021; Zaslavsky et al., 2021) and it generated precise quantitative predictions that can be tested in our case.

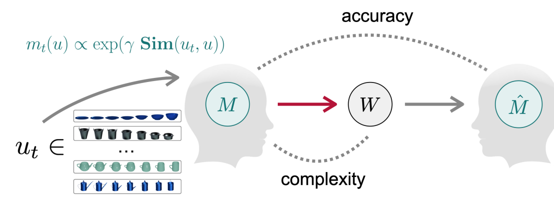

Figure 3: The communication model of Zaslavsky et al. (2018), which is grounded in an underlying similarity-based representation of the domain. See main text for details.

This framework is based on the communication model shown in Figure 3, in which a speaker is modeled as a stochastic encoder $q(w|m)$ probabilistically mapping meanings m to labels w, and a listener as an ideal observer mapping from labels to inferred meanings $\hat{m}_w$. For each language group and for each continuum, we estimate an aggregate speaker as the empirical labeling distribution from the 2AFC tasks described in the previous section.

The speaker’s meanings, m, are assumed to be belief states, formally defined as a probability distribution $m(u)$ over elements in a domain $\mathcal{U}$. We will therefore consider a model instantiation for each of the six continua from the previous section, taking $\mathcal{U}$ to be the 7 objects (or steps) in the continuum. As explained in the following subsection, we instantiate these six models by grounding meanings in non-linguistic similarity spaces. This grounding is the only component that varies across the six model instantiations, and importantly, the formal notion of efficiency remains exactly the same.

According to the IB framework, systems are efficient to the extent that they satisfy a tradeoff between minimizing complexity—measured in bits as the resources needed to compress a thought into a word—and minimizing distortion of meanings, which reflects the listener’s error in reconstructing the speaker’s intended meaning. More formally, the complexity of any given encoder (or empirical labeling system) is defined by the mutual information it induces between meanings and labels,

$$I_q(M;W) = \sum_m p(m)q(w|m)\log\frac{q(w|m)}{q(w)}, \quad (1)$$

where $q(w) = \sum_m p(m)q(w|m)$ is the marginal distribution of labels, and $p(m)$ denotes a distribution over steps in the continua which, for simplicity, we take here to be uniform. Accuracy depends on how well a listener can reconstruct the speaker’s intended meaning, formally defined in terms of the Kullback-Leibler (KL) divergence between speaker and listener meanings,

$$\mathbb{E}_q\left[D_{\mathrm{KL}}\left[M\|\hat{M}\right]\right] = \sum_{m,w} p(m)q(w|m)D\left[m\|\hat{m}_w\right]. \quad (2)$$

Minimizing Eq. 2 amounts to maximizing $I(W;U)$ (Tishby et al., 1999; Zaslavsky, 2020), which can be interpreted as

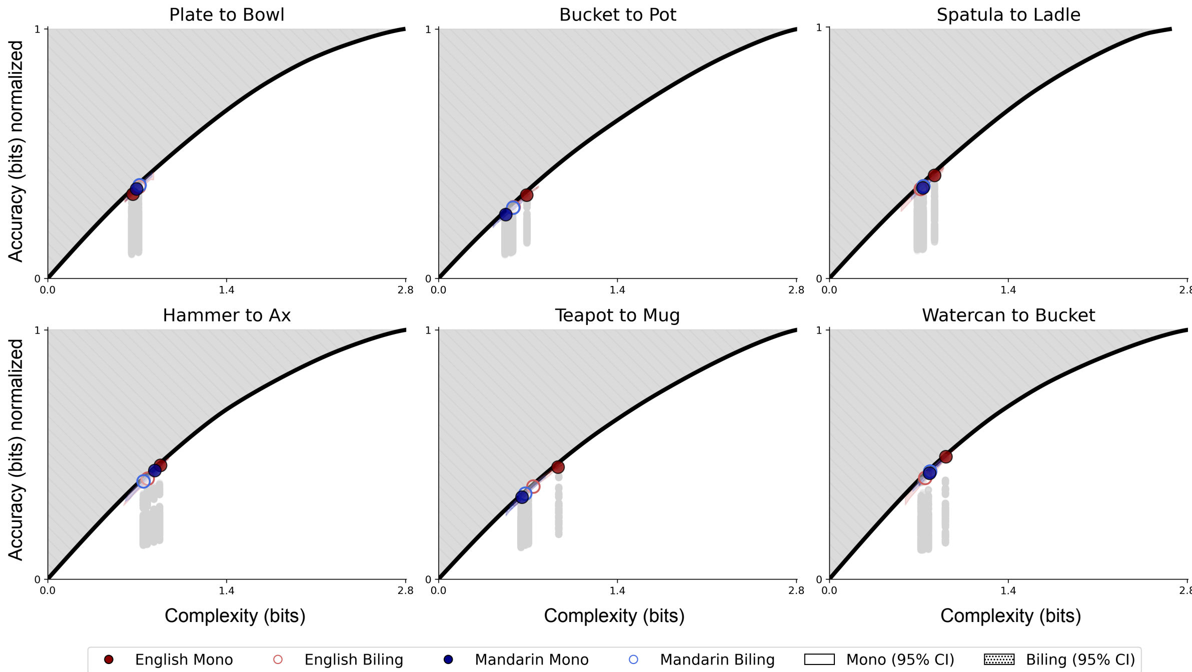

Figure 4: Complexity-accuracy tradeoffs for both monolinguals and bilinguals across the six continua. Colored dots and colored regions mark the means and 95% CIs, respectively, of complexity and accuracy of each language group based on 1000 bootstrapped resamples from the participant population. Gray dots mark all 5040 permutations of each actual system.

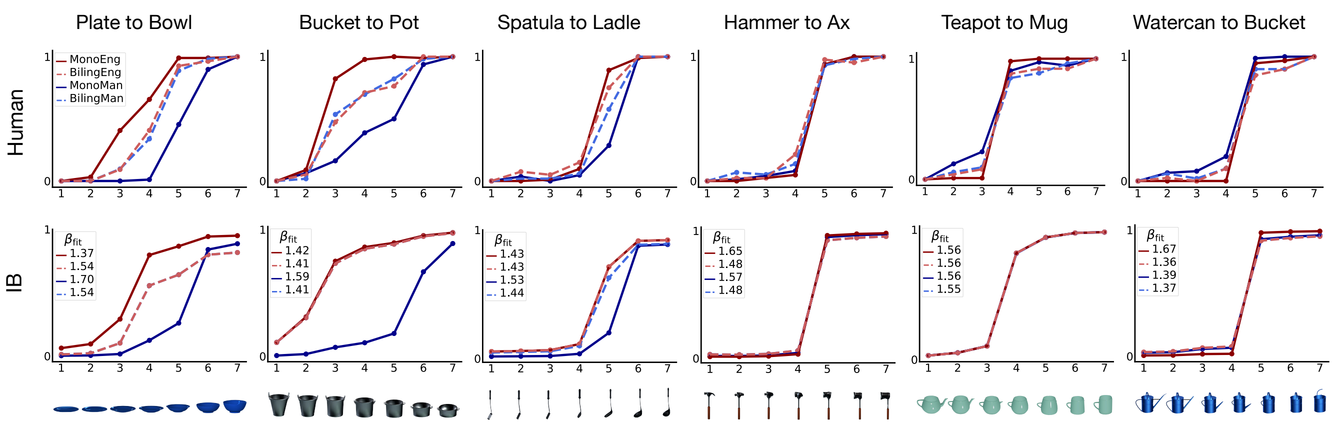

Figure 5: Comparison of human (top, from Figure 2a) and fitted IB (bottom) category boundaries, with corresponding β values.

the information-theoretic accuracy of the lexicon. Taken together, a category system that is optimally efficient minimizes the overall IB objective function:

$$\mathcal{F}_{\beta}[q(w|m)] = I_q(M;W) - \beta I_q(W;U), \tag{3}$$

where $\beta \geq 1$ is a tradeoff parameter. The solutions to this optimization problem at different values of β define the IB theoretical limit of efficiency (see Figure 4, black curves), which is the upper bound on achievable complexity-accuracy tradeoffs. If both monolinguals and bilinguals operate similarly under pressure to maintain efficient category systems, then their systems should lie near this bound, possibly with different tradeoffs β as their systems do vary. However, if bilinguals operate under high demands from other (e.g., sociolinguistic) constraints that compete with efficiency, they may be pulled away from the bound relative to monolinguals. To test which of these possibilities are consistent with our data, we first need to compute the bound for each continuum, which requires specifying a grounded meaning space.

Grounding meanings in non-linguistic similarities

For each of the six continua under study, we model the speakers' meanings $m(u)$ as probabilistic mental representations of the continua. Following prior work (e.g., Regier, Kemp,

& Kay, 2015; Xu et al., 2016), we define the meaning of each target referent $u_t \in \mathcal{U}$ as a similarity-based distribution, $m_t(u) \propto \exp(\gamma \mathrm{Sim}(u_t, u))$. The value of γ determines the degree of speaker uncertainty, which we approximate using the domain-agnostic method from (Eisape, Levy, Tenenbaum, & Zaslavsky, 2020; Zaslavsky et al., 2021). $\mathrm{Sim}(u_t, u)$ denotes the similarity between u_t and u, which we estimate empirically by collecting pairwise similarity rating using the non-linguistic task described below. Because prior work has shown that similarity rating judgments are highly correlated across languages (Ameel et al., 2005; Malt, Sloman, Gennari, Shi, & Wang, 1999; Malt et al., 2003; Maldonado, Zaslavsky, & Culbertson, 2023), we focus here on similarity data from monolingual English speaking participants.

Participants. We recruited 40 monolingual English speakers residing in the USA (22 females, 18 males; $M_{\text{age}} = 41.5, SD = 14.8$) online, excluding those with second-language proficiency score above 3 out of 7.

Procedure. Participants rated the similarity of all pairwise step combinations (0 = "Not at all similar" to 4 = "Same"). Ratings were averaged across participants to create a similarity matrix for each continuum.

Bilinguals are near-optimal, akin to monolinguals

Given the IB model instantiation for each of the six continua, we computed the complexity and accuracy for each language group and compared these values to the IB theoretical limit of efficiency. Figure 4 shows that across all six continua, the category systems of all language groups—English and Mandarin monolinguals and bilinguals in either language—lie extremely close to the theoretical bound. To account for estimation errors in the empirical complexity-accuracy tradeoffs, we report bootstrapped average values (colored dots) and 95% confidence intervals (colored regions) based on 1000 resamples (Figure 4). All variation in our estimates is confined to a narrow region near the bound, with populations differing only in their precise location along it. This means that while we cannot draw conclusions as to the relative position of each language group along the bound, we can conclude with confidence that they are all near-optimal. To further test this for significance, we considered all possible permutations of each language group's empirical category system, i.e., $p(w|m)$, within each continuum (yielding $7! = 5040$ counterparts per group per continuum), and assessed the deviation from optimality as the test statistic.[1] We found that all groups across all continua are significantly efficient ($p < 10^{-5}$).

Importantly, our finding that both monolinguals and bilinguals are highly efficient holds both in the 'same continua' and 'different continua' domains. This suggests that while bilinguals exhibit semantic convergence in their categories, this convergence may still be constrained by pressure to maintain IB-efficient category systems independently of other factors, such as sociolinguistic pressures, that are likely to affect bilinguals in different ways compared to monolinguals.

Finally, we note that the qualitative behavior of the optimal IB systems captures some aspects of the behavioral data, but also deviates from it in interesting ways. First, in contrast to the 2AFC task, the model is not restricted to using only two labels. In fact, shifts in category boundaries, as observed in the behavioral data, tend to emerge in the model from competition between more than two categories. This raises an interesting prediction that perhaps bilinguals may have intermediate categories in addition to the two endpoint labels. Another source for boundary shifts in the model could stem from our assumption of a uniform prior, whereas a non-uniform one could shift boundaries by weighting some meanings more. These two observations about the model suggest the need for theory-driven experiments in future work, considering free-labeling and prior elicitation tasks. Second, as can be seen in Figure 5, while the model captures the category boundaries in the 'same continua,' the qualitative fit in the 'different continua' is somewhat mixed. We fit each empirical system in our dataset with an IB-optimal encoder by first reducing the encoder to the modal categories of step 1 and 7, and then selecting the encoder that minimizes the squared error between the probabilities it predicts for these two categories and the average empirical probabilities from Figure 2. The main discrepancy of the model is in the 'Bucket to Pot' continuum, where it fails to account for the bilingual convergence, in contrast to the other two 'different continua' where the model can capture the observed convergence.

Discussion

In this work we addressed two open questions related to categorization in bilinguals: (1) How do they integrate non-overlapping category systems from different languages? And (2) to what extent do they maintain efficient category systems? By combining theory-driven experiments and an information-theoretic efficiency analysis, we showed that bilinguals exhibit semantic convergence in their category systems, departing from monolinguals in this respect, but at the same time their category systems remain near-optimally efficient, akin to monolinguals. These findings provide new insight into how bilinguals may navigate cross-linguistic semantic variation and suggest that a drive for efficiency is a key factor shaping bilinguals, in addition to other, e.g., sociolinguistic, factors. Our work also underscores the need for further theory-driven research to tease apart potential computational accounts for the observed shifts in the bilingual category boundaries, such as differences in priors or competition between categories. Finally, another important direction for future research is to study bilinguals from different sociolinguistic backgrounds. This would help determine whether the merged category boundary is always shaped equally by both languages or if its position shifts based on how often a person is exposed to each language's boundaries.

[1] We also repeated this analysis with a measure of similarity between the full probabilistic structure of the optimal and attested systems and found similar results.

References

Ameel, E., Malt, B. C., Storms, G., & Van Assche, F. (2009). Semantic convergence in the bilingual lexicon. *Journal of memory and language*, *60*(2), 270–290.

Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of memory and language*, *53*(1), 60–80.

Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. University of California Press.

Blanco-Elorrieta, E., & Caramazza, A. (2021). A common selection mechanism at each linguistic level in bilingual and monolingual language production. *Cognition*, *213*, 104625.

Eisape, T., Levy, R., Tenenbaum, J. B., & Zaslavsky, N. (2020). Toward human-like object naming in artificial neural systems. In *Bridging AI and Cognitive Science workshop at the International Conference on Representation Learning (ICLR).*

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407. doi: 10.1016/j.tics.2019.02.003

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, 1–23. doi: 10.1146/annurev-linguistics-011817-045406

Maldonado, M., Zaslavsky, N., & Culbertson, J. (2023). Evidence for a language-independent conceptual representation of pronominal referents. In *Proceedings of the 45th annual meeting of the cognitive science society.*

Malt, B. C. (1995). Category Coherence in Cross-Cultural Perspective. *Cognitive Psychology*, *29*(2), 85–148. doi: 10.1006/cogp.1995.1013

Malt, B. C. (2024). Representing the World in Language and Thought. *Topics in Cognitive Science*, *16*(1), 6–24. doi: 10.1111/tops.12719

Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *WIREs Cognitive Science*, *4*(6), 583–597. doi: 10.1002/wcs.1251

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262.

Malt, B. C., Sloman, S. A., & Gennari, S. P. (2003). Universality and language specificity in object naming. *Journal of memory and language*, *49*(1), 20–42.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. doi: 10.1037/0033-295X.85.3.207

Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., & Kemp, C. (2021). The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, *118*(49), e2025993118. doi: 10.1073/pnas.2025993118

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication Informativeness and simplicity as competing principles. In B. MacWhinney & W. O'Grady (Eds.), *The Handbook of Language Emergence* (pp. 237–263). Wiley.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573–605.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottlneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 368–377.

Van Hell, J. G., & De Groot, A. M. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and cognition*, *1*(3), 193–211.

White, A., Malt, B. C., & Storms, G. (2017). Convergence in the bilingual lexicon: A pre-registered replication of previous studies. *Frontiers in psychology*, *7*, 2081.

White, A., Storms, G., Malt, B. C., & Verheyen, S. (2018). Mind the generation gap: Differences between young and old in everyday lexical categories. *Journal of Memory and Language*, *98*, 12–25.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical Semantic Chaining and Efficient Communication: The Case of Container Names. *Cognitive Science*, *40*(8), 2081–2094. doi: 10.1111/cogs.12312

Zaslavsky, N. (2020). *Information-Theoretic Principles in the Evolution of Semantic Systems*. Unpublished doctoral dissertation, The Hebrew University of Jerusalem.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942. doi: 10.1073/pnas.1800521115

Zaslavsky, N., Maldonado, M., & Culbertson, J. (2021). Let's talk (efficiently) about us: Person systems achieve near-Optimal compression. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society.*

Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. In *41st Annual Meeting of the Cognitive Science Society.*