Found in translation

Oppenheim, Gary; Wu, Yan Jing; Thierry, Guillaume

Cognitive Science

DOI:
10.1111/cogs.12618

Published: 01/07/2018

Peer reviewed version

Found in translation: Late bilinguals do automatically activate their native language when they are not using it

Gary Oppenheim[1,2], Yan Jing Wu[3], and Guillaume Thierry[1,4]

[1] School of Psychology, Bangor University, UK
[2] Rice University, Houston, TX, USA
[3] Faculty of Foreign Languages, Ningbo University, China
[4] Centre for Research on Bilingualism, Bangor University UK

Abstract

In their paper "Do bilinguals automatically activate their native language when they are not using it?", Costa, Pannunzi, Deco, and Pickering (*Cognitive Science*, 2016) proposed a reinterpretation of Thierry and Wu's (2004, 2007) finding of native language-based (Chinese, L1) ERP effects when they tested Chinese-English late bilinguals exclusively in their second language (English, L2). Using simulations in a 6-node Hebbian learning model (three L1 nodes, three L2 nodes), Costa and colleagues suggested that form overlaps in L1 between otherwise unrelated words create a persistent relationship between their L2 translations. In this scenario, words in the nascent L2 lexicon overlapping in their L1 translations would become linked during learning because of the form overlap in L1; once the L2 words are learned, the direct link between them would be sufficient to generate robust apparently 'L1-mediated' priming without requiring any activation of L1 translations. Costa et al. contend that links between L2 words remain beyond the learning phase, even after links to L1 representations have been severed, and thus that their model affords an alternative account to (but not a rebuttal of) Thierry & Wu's claim of language non-selective activation –or automatic activation of translation equivalents– in late bilinguals. In this response, we build on Costa et al.'s original simulation code, showing that it can only reproduce L1-independent priming when implementing the L1 disconnection in their particular way. By contrast, when severing inter-language connections bi-directionally, their model fails to retain any sizeable influence of L1 form overlap on L2 activations. The model is not the theory, however, and we discuss several issues that would need to be addressed in further attempts to model language non-selective activation in late bilinguals.

Keywords
Language co-activation, event-related potentials, modelling, learning, translation equivalent

Corresponding author
Guillaume Thierry
School of Psychology
Bangor University
LL57 2AS Bangor, Wales, UK
g.thierry@bangor.ac.uk

Introduction

When bilinguals use their second language, are they able to completely avoid accessing their native language? After decades of reports of cross-language effects in bilinguals (e.g., De Groot et a., 1991; Brysbaert, 1998; Van Hell and Dijkstra, 2002; Duyck, 2005; Duyck et al., 2007), Thierry and Wu (2004, 2007) demonstrated that late Chinese-English bilinguals automatically and unconsciously activate native translation equivalents when reading L2 English words using event-related potentials (ERPs) in an experimental context entirely uncontaminated by L1 Chinese. On each trial participants read or heard two English words presented in a sequence, like 'train' and 'ham', and reported whether or not these words were related in meaning. Half of the word pairs had a hidden relationship through Chinese translations: For instance, 'train' and 'ham' translate into *huo che* and *huo tui*, in Mandarin Chinese, and thus share their first syllable. Although bilingual participants failed to consciously detect this link and although no sign of it was found in their reaction times or accuracy data, ERP data revealed the priming effect that one would expect if they had accessed the translation of the English words in their native language Chinese. Thierry and Wu (2004, 2007) thus concluded that late bilinguals automatically and unconsciously access L1 translation equivalents when processing L2 words (see also Wu and Thierry, 2010a; 2012a, 2012b; Spalek et al., 2012; Wu et al., 2013).

In a paper built around a Hebbian learning model, Costa et al. (2016) recently suggested an alternative explanation: rather than revealing ongoing contributions of L1 to L2 processing, Thierry & Wu's findings may simply reflect a particular way that L1 shaped L2 at a much earlier stage during the L2 learning process. Given that *huo che* and *huo tui* sound the same in Mandarin, then maybe their lexical representations will tend to be activated at the same time in Mandarin, and because neurons that fire together wire together, *huo che* and *hui tui* will become linked. And then their respective translations in English, 'train' and 'ham', will also end up becoming linked to one another, within L2. Thus, one might not actually need to access L1 to get L2 effects that *look like* they arise from L1 overlap.

Costa et al.'s (2016) proposal offers an intriguing alternative explanation for Thierry and Wu's findings, and the choice to implement it as a computational model brings a laudable rigor to their approach. We were curious about some of Costa et al.'s modelling decisions and how much the simulated results may depend on such implementation details. We also wondered how their approach might be extended to better approximate experimental tasks used when testing human participants, such as semantic judgements. Therefore, we wrote to the authors who kindly shared with us the original code they used for their simulation. The present paper deals with three issues relating to Costa et al.'s approach: (1) we provide a detailed account and small extension of Costa et al.'s model, finding that, when L1 and L2 are actually bidirectionally disconnected, it fails to provide a convincing alternative for our empirically observed phenomenon; (2) we discuss some limitations of Costa et al.'s approach to simulating bilingual language development (abrupt disruption of the links between L1 and L2, lack of consideration for unlearning); (3) we introduce three further conceptual issues that are important to situating any such model in the wider context of bilingual language use (what we know about bilingual lexical access, the communicative

function of language, and bilingual diversity). Even though learning a second language may lead to links existing between L1 lexical forms to be inherited by L2 representations, language non-selective access (i.e., activation of native language translation equivalents) in late bilinguals remains the most parsimonious account of what happens in the mind of late (Chinese-English) bilinguals when they are exposed to second language words, whether spoken or written.

1. The model as conceived and tested by Costa et al.

1.1. Architecture

The model contains six nodes, each representing one word. Three nodes represent words in L1 {$train_M$, ham_M, $apple_M$} and three represent words in L2 {$train_E$, ham_E, $apple_E$}. Every node is connected to every other node, with a weight initialized as 0. To represent L1 phonological overlap between $train_M$ and ham_M, however, the weight of the connection between them is initialized as c_{Ph} (see **Table 1** for parameter descriptions). Following a standard convention in Hebbian networks, where nodes are assumed not to carry over activation between timesteps, connections from each node to itself are maintained at zero.

Table 1 - Parameters of Costa et al.'s model[1]

Parameter	Implemented value	Description
max(τ)	30000	Number of training trials
Ω	6000	Asymptotic scaling constant (learning drops off as a function of 2/(1+exp(τ /Ω)))
ν_H	40	Input activation for the L2 target word
ν_{H2}	15	Input activation for the L1 translation word
ν_{VL}	4	Input activation for all other words
$\Delta\nu$	2	Variability in input activation
c_{Ph}	0.12	Connection weight imposed between phonologically similar words in L1
α_L	0.001	Learning rate
Θ_L	20	Threshold for learning and the mean of a sigmoidal function to scale weight changes
β_L	6	Standard deviation for a sigmoidal function to scale weight changes

[1] The values listed here are those effectively implemented in the simulation of Costa et al., see Costa, A., Pannunzi, M., Deco, G. and Pickering, M. J. (2017), Corrigendum for: Do bilinguals automatically activate their native language when they are not using it?. Cogn Sci. doi:10.1111/cogs.12577

1.2. Activation

Activation begins with a vector of external inputs to nodes {$train_M$, ham_M, $apple_M$, $train_E$, ham_E, $apple_E$}, with values reflecting the parameters {ν_H, ν_{H2}, ν_{VL}}. For example, when cuing the English word train, the nodes {$train_M$, ham_M, $apple_M$, $train_E$, ham_E, $apple_E$} would take input values of {ν_{H2}, ν_{VL}, ν_{VL}, ν_H, ν_{VL}, ν_{VL}}, respectively. On each trial, this input vector is increased by a vector of uniformly distributed random values, with values between 0 and $\Delta\nu$; the initial activation of each node is thus simply its noisy net input. Thereafter, the initial activation of each node *i* at each time step *t* is assumed to be the sum of all of the other nodes *j* in the network at the previous time step, times the weight of their connections to that node w_{ij}.

$$a_{i_t} = \sum w_{ij} a_{j_{t-1}}$$

Activation then continues to spread until the network settles into a stable state. For any vector of initial activations, A_{t0}, and matrix of weights, W, this stable state $A_{t\infty}$ is assumed to be approximated by minimizing the error for an implied set of equations:

$$A_{t_\infty} = lsqr((W - [identity\ matrix]), -A)$$

1.3. Learning

For each trial, τ in 1: max(τ), after the network has settled, the weights of all connections w_{ij}, into each node *i* that has exceeded an activation threshold, Θ_L, are modified according to the following equation:

$$w_{ij_{\tau+1}} = w_{ij_\tau} + \left(x{\sim}U(0, \alpha_L)\right) \times normcdf\left(a_{j_\tau}, \mu = \Theta_L, \sigma = \beta_L\right) \times \left(1 - w_{ij_\tau}\right) \times \left(\frac{2}{1 + e^{\frac{\tau}{\Omega}}}\right)$$

The first component of this equation specifies that the old weight should serve as the basis for the new weight. The next component specifies that the amount of the weight change should be scaled according to a uniformly distributed random number in the range [0, α_L]. The next component specifies that the amount of the weight change should be scaled according to a [0,1] sigmoidal transformation of the sending node's activation (implemented as the probability of observing a value less than a_j in a normal distribution with mean Θ_L and standard deviation β_L, *e.g.*, resulting in a value of .5 when $a_j = \Theta_L$, and .9772 when $a_i = \Theta_L + 2*\beta_L$). The next component specifies that any weight should asymptote around a maximum value of 1. The final component specifies that weight changes should be scaled such that they will be very large when training begins, before asymptoting toward zero (e.g., for $\Omega = 6000$, on training trials $\tau = \{1, 10000, 20000, 30000\}$, this component would scale any weight change by {.9999, .3177, .0689, .0134}); assuming that the 30,000 training trials represent the Thierry & Wu's participants' pre-experiment exposure to English, this component implements an assumption that they would learn new English translations of

Mandarin words 100 times more slowly as college students than they did when beginning to learn English as children.

Thus, the learning rule specifies that weight changes should only occur when the receiving node is active above some threshold, but also be greater when the sending node is more active. Although the learning rule allows for randomness in the extent of a weight increase, it nonetheless specifies that connection weights can never decrease; any temporary co-activation of two nodes is assumed to form a connection between them that will persist forever. The rule also specifies that connection strengths have some upper limit beyond which they cannot remain, perhaps reflecting a biological saturation point. And finally, the rule further specifies that weight changes must decelerate as the system ages, enforcing something like a 'critical period'.

1.4. Simulations

1.4.1. Training. Exactly replicating Costa et al.'s implementation, the network was initialized as specified above, and trained for 30,000 trials. On each trial, one L2 word was randomly selected as the target (thus each L2 word was trained approximately 10,000 times); its node received ν_{H} units of input activation, its L1 translation received ν_{H2} units of activation, and all other nodes received ν_{VL} units of input activation. The settling state of the network was then estimated as described in 1.2, and connections were modified as described in 1.3, producing a trained network with connections as depicted in **Fig. 1a**.

1.4.2. L1-intact simulation (Fig. 1a). Replicating Costa et al.'s L1-intact simulation, as the network's 30,001st trial, one L2 word was randomly selected as the target and activated just as previously specified for training: its node received ν_{H} units of input activation, its L1 translation again received ν_{H2} units of activation, and all other nodes received ν_{VL} units of input activation. The settling state of the network was then estimated as described above, with the resulting activation of each node recorded (as the basis for Costa et al.'s Figure2b), but no further weight changes were applied. This testing procedure was applied 8000 times, so each L2 word was tested approximately 2,667 times.

1.4.3. L1-'disabled' simulation (Fig. 1b). Replicating Costa et al.'s L1-disabled simulation, first, all connections to L1 nodes from L2 nodes (L2 → L1) were set to 0, implementing a restriction such that an L2 word cannot directly activate its L1 translation, nor any other L1 word. Critically, however, in Costa et al.'s L1-disabled simulation, L1 words continued to be fully activated and all connections to L2 from L1 (L1 → L2) remained intact (i.e. not set to 0); we worry that some readers may have misunderstood these important details when Costa et al. described their simulation as having "removed all L1 representations" (p12) and "restricted activation to only one language" (p13), thereby "removing any on-line influences between L1 and L2," (p7) and having effectively "turned the model monolingual" (p13).

Then the testing procedure continued exactly as in the L1-intact simulation and Costa et al.'s L1-disabled simulation. As the network's 30,001st trial, one L2 word was randomly selected as the target and activated just as previously specified for training: its node received ν_{H} units

of input activation, its L1 translation again received ν_{H2} units of activation, and all other nodes received ν_{VL} units of input activation. The settling state of the network was then estimated as described above, with the resulting activation of each node recorded (as the basis for Costa et al.'s Figure 2b), but no further weight changes were applied. This testing procedure was applied 8000 times, so each L2 word was tested approximately 2,667 times.

1.4.4. Our L1-disabled simulation (Fig. 1c). First, all L2 → L1 connections were set to 0, implementing a restriction that an L2 word cannot directly activate its L1 translation, nor any other L1 word. In our simulation, though, all L1 → L2 connections were also set to 0, implementing a corresponding restriction that an L1 word cannot directly activate its L2 translation either (nor any other L2 word).

Then the testing procedure continued exactly as in Costa et al.'s L1-disabled simulation, except that in our simulation the L1 translation received only ν_{VL} units of input activation (on the assumption that ν_{VL} represents a resting level of activation for all nodes) instead of the full ν_{H2} units of activation applied in Costa et al.'s L1-disabled simulation. Thus, as the network's 30,001st trial, one L2 word was randomly selected as the target and activated: its node received ν_{H} units of input activation, and all other nodes (including its L1 translation) received ν_{VL} units of input activation. The settling state of the network was then estimated as described above, with the resulting activation of each node recorded (as the basis for Costa et al.'s Figure 2b), but no further weight changes were applied. This testing procedure was applied 8000 times, so each L2 word was tested approximately 2,667 times.

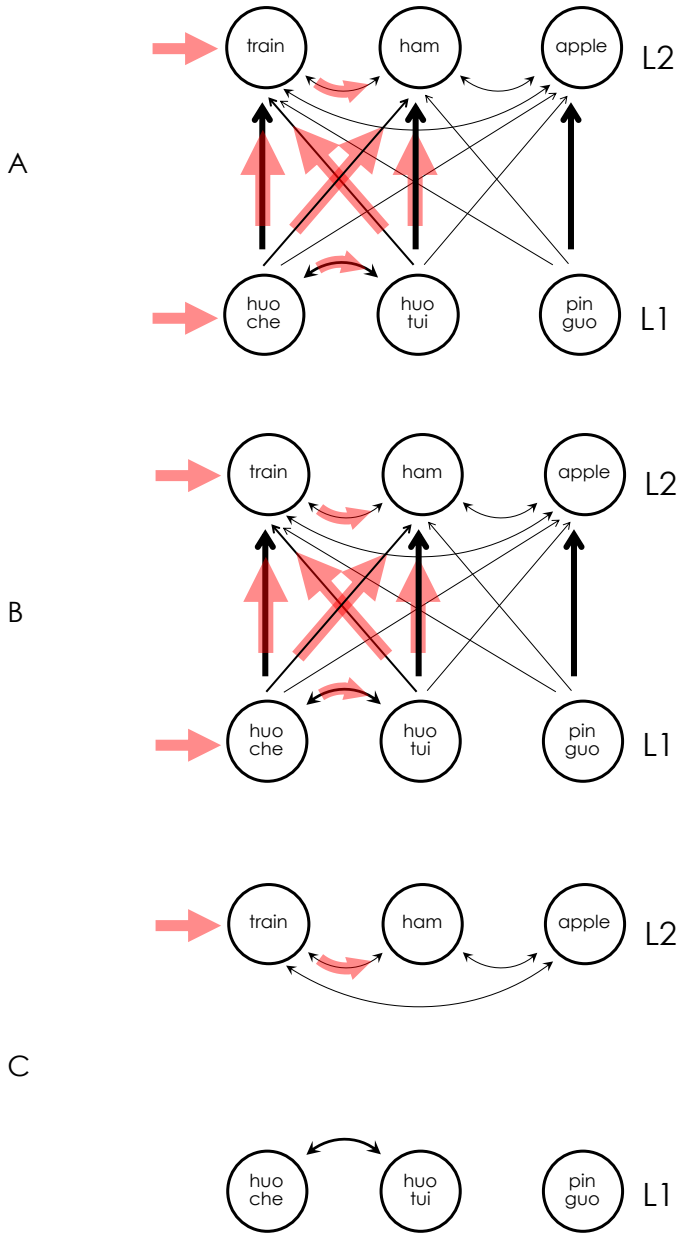

Figure 1 – Schematic of the cross-linguistic pattern-completion model as implemented in Costa et al.'s L1-connected simulation (A), their L1-disconnected simulations (B), and our L1-fully-disconnected simulation (C). Line thickness approximates connection strength on a logarithmic scale, completely omitting connections of weight 0. Overlays depict activation flow through major connections supporting 'ham' over 'apple'. Panels A and B are identical because Costa et al.'s L1 disconnection only removed connections from L2 to L1, which never actually developed in the first place.

1.5. Results

Results of our simulations are presented in **Figure 2**. As they should, our Panels A-D closely resemble Panels A-D in Costa et al. 's Figure 2. They are, after all, generated from the same code, using the same set of parameters. As in Costa et al. 's simulations, they demonstrate that the model learns its connections through experience (**Fig. 2a)**, and although the model does not associate the English words 'train' and 'ham' when it is first initialized (**Fig. 2b**; Cohen's $d < 0.01$ for the difference between 'ham' and 'apple' activations), its experience-

driven learning leads it to activate the English word ‘ham’ more than the English word ‘apple’ when prompted by the English word ‘train’ (**Fig. 2c**; Cohen’s *d* = 0.37). As illustrated in **Figure 1a**, activating ‘train’ and ‘huo che’ in the L1-intact simulation activates ‘ham’ via a set of L1-dependent pathways (‘huo che’ → ‘huo tui’ → ‘ham’, ‘huo che’ → ‘ham’, and ‘huo che’ → ‘train’ → ‘ham’) as well as an L1-independent pathway (‘train’ → ‘ham’). And as in Costa et al.’s L1-disabled simulation, these same pathways (**Fig. 1b**) continue to activate the English word ‘ham’ when activating the English word ‘train’ and its Mandarin translation, despite the connection from the English word ‘train’ to its Mandarin translation having been severed (**Fig. 2d**; Cohen’s *d* = 0.37).

However, Costa et al.’s central claim was not merely that learned associations cause the English words ‘train’ and ‘ham’ to become co-activated, but rather that such co-activation could emerge entirely from learned associations within English, and thus that the empirical observation of a co-activation would not constitute evidence that late bilinguals activate their L1 when processing words in their L2. Therefore, the critical test of Costa et al.’s theoretical claim is a test that they did not conduct: What happens to the English word ‘ham’ when we activate the English word ‘train’, but, as illustrated in **Figure 1c**, we (1) refrain from activating its Mandarin translation and, crucially, (2) sever connections both to L1 from L2 and to L2 from L1? **Fig. 2e** shows that, under these conditions, the phenomenon that Thierry & Wu claimed as evidence for necessary L1 activation essentially disappears (Cohen’s *d* = 0.046). Without activating L1, and without allowing L1 to activate L2 and L2 to activate L1, the difference between the activations of the English words ‘ham’ and ‘apple’ is quite minimal, even in a model that was specifically designed to produce such a difference.

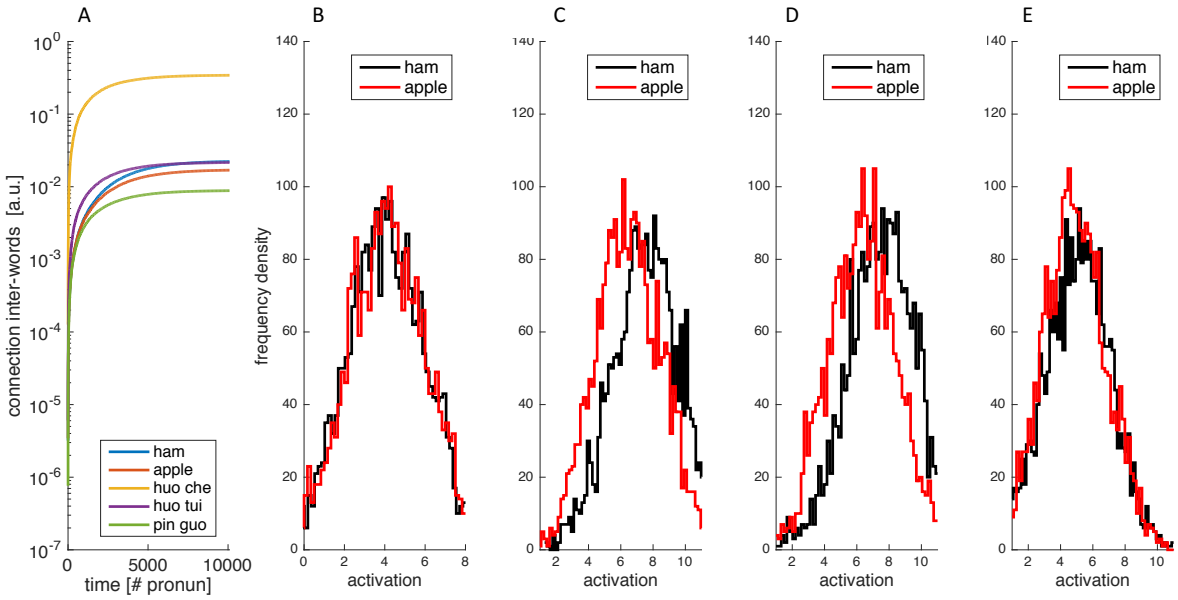

Figure 2 – Simulation results. (a)-(d) replicate results those reported by Costa et al. (2016). (a) Connection strength to train grows over 10,000 training trials. (b) Pre-training L2 lexical activation. (c) Post-training L2 lexical activation. (d) Post-training L2 lexical activation, after severing just L2→L1 connections, as in Costa et al.’s L1-“disabled” simulation. (e) Post-training L2 lexical activation, after severing *all* between-language connections, as in our L1-disabled simulation.

One may argue that our L1-disabled simulation still shows some hint of an effect: in **Figure 2E**, the mean activation of ham (4.98) is about 3% larger than the mean activation of apple (4.82). But this difference is very weak, and if Costa et al.'s model allows for reasonable signal-to-noise ratios, that difference could hardly have enabled several replications of the effect (e.g., Wu & Thierry, 2010a; Wu & Thierry, 2012a; 2012b; Wu et al., 2013). Thus, although a fully disconnected implementation of Costa et al.'s model can lead to lasting L1-based associations between L2 words without persistent L1 activation, any such L1 associations would be far subtler than its proponents have claimed.

2. Suggestions for simulation improvements

The model is not the theory, however, and it is possible that slight adjustments of Costa et al.'s model, or a different implementation of the same basic ideas would generate stronger L2-based effects of L1-derived associations. Therefore, it is important to consider some of the model's motivating ideas, which it would necessarily share with any alternative implementation.

2.1. Incorporating mechanisms for unlearning

One curious aspect of Costa et al.'s simulation is that the model stopped learning immediately after disconnecting L1 from L2. What would happen if the model continued to modify its connections based on experience in its L1-disconnected state? Would the claimed connections within L1 persist, or would they fade away? The answer appears to rest on aspects of the model that lack both strong theoretical and computational support. Most important among these is the model's implementation of Hebbian weight increases as its sole basis for synaptic changes. Hebbian learning is a family of approaches that are elegant, neurally plausible, work on strictly local information, and have proven extraordinarily successful at unsupervised associative learning. Many recent advances in machine learning, for instance, employ learning rules that can be considered part of the Hebbian family. As such, the selection of *some* Hebbian learning rule may seem like a simple, theory-neutral approach, even though it actually implements a very strong assumption: the model can only learn through excitatory connections, only ever strengthening them. Elsewhere, mechanisms for reducing association weights are common to most computational modelling frameworks (e.g. Rescorla & Wagner, 1972; Rumelhart & McClelland, 1986; Oppenheim, Dell & Schwartz, 2010), and researchers have long recognized their value in Hebbian learning systems (e.g. Hopfield, 1982), specifically increasing a system's stable storage capacity. Such association reduction can be implemented in many ways (e.g. unlearning, weight decay, weight normalization), but all solve the same basic problem arising from the Hebbian weight-strengthening rule: whenever two patterns overlap, a Hebbian network queried with part of one will complete both patterns. If any node is erroneously activated, continued association learning will add that node to the pattern, snowballing until any node will activate the entire network, thus catastrophically failing to recover any individual pattern (in biological networks, such activation might correspond to a seizure).

In fact, Costa et al.'s simulation of learning of L1-mediated associations within L2 provides a snapshot of such a snowballing effect in progress; only its hard-coded critical period function prevents its 30,000 training trials from resulting in a superposition catastrophe, where it would be unable to distinguish between train and ham[2]. Like other Hebbian systems, Costa et al.'s model could thus benefit from incorporating some mechanism for unlearning, but doing so would entirely eliminate its account of Thierry & Wu's data, because such functions would stabilize the network by unlearning precisely the kinds of spurious associations that underlie the account. Thus, omitting of a mechanism for unlearning may seem like a mere simplification, but it is in fact crucial to Costa and colleagues' theoretical account, and any similar account would need to make the same omission.

2.2. Incorporating more gradual L1 disconnection

It is difficult to conceptualise a real life equivalent of Costa et al.'s abrupt testing method — suddenly dismissing connections from L2 to L1 and probing the state of the obliterated network. Although lesioning a connection is a common way to assess its contributions, Costa et al.'s proposal goes further, claiming not only that L2-internal links may contribute to L1-based effects, but that their inter-lingual lesioned network may actually represent an unimpaired late bilingual's normal state. At the least, it seems rather implausible that a simulation abruptly lesioning L1 representations can provide a realistic account of unimpaired bilinguals' transition from an earlier to a later stage of L2 acquisition.

Note that the need for more gradual L1 disconnection is not a mere implementational detail, because it should clearly interact with any implemented unlearning mechanism, reducing and eliminating Costa et al's hypothesised within-L2 connections. In fact, Costa et al. themselves suggest that, if the connections underlying their effect are "not refreshed regularly, [they] may disappear via [an unimplemented mechanism for] depotentiation" (p15). Thus, although both their and our simulations suggest that a within-L2 basis for the effect may briefly persist after abruptly severing L1↔ L2 connections, even according to Costa et al.'s own account, it requires regular 'refreshing' via normally intact L1↔ L2 connections. As L1↔ L2 connections gradually weaken, they thereby lose their ability to refresh hypothesised within-L2 connections via their stated Hebbian learning mechanisms, and thus reduce their ability to account for Thierry & Wu's empirical findings. In other words, even according to Costa et al.'s own internal logic, one can only account for L1-based effects without L1↔ L2 connections if those connections are usually intact.

3. Broader conceptual issues

3.1. Post-hoc ad-hoc modelling

Although model-building rarely happens in a vacuum, without consideration of empirical data, there is a process that distinguishes theory-driven cognitive modelling from results-

[2] this can be verified by e.g., increasing ω from 6000 to 15000, or by simply running the model with its originally published set of parameters.

driven AI. The first step aims to characterise the goal of a process (Marr's, 1981, "computational theory" level), and this is perhaps why cognitive models usually focus characterising performance on particular tasks with well-defined inputs and outputs. Costa et al's approach clearly fails to define a task (simulating effects during semantic judgements without a semantic layer; more on this below), and it is not immediately clear what challenge the modelled system might be trying solve by storing direct excitatory links between unrelated words within a language.

The second step is to implement a guess about specific architectures and algorithms that the mind might use to achieve its processing goal. For instance, if we assume the model's goal is to associate strongly activated L2 words with their strongly activated L1 translations, using Hebbian learning to build these associations might be a reasonable approach. What is not clear is why such association-building processes should also apply to weakly[3] co-activated words within a language, or why such associations should always increase and never decrease. One might argue that there is considerable evidence for within-language excitatory lexical associations, but that argument misses a crucial point: those within-language lexical associations tend to be based on distributional information and transitional probabilities, as might be modelled using a simple recurrent network (e.g. Elman, 1991; Chang et al., 2006) to predict upcoming words in a sequence. The kind of within-language lexical association that Costa et al. claim, on the other hand, seems to exist only to artificially account for Thierry and Wu's data.

3.2. Compatibility with empirical data

In Costa et al.'s original model, after L2 is fully acquired, L2-to-L1 connections are no longer available whilst L1-to-L2 connections remain active. This would imply either (a) unbalanced connections between two lexica, with stronger links from L1-to-L2 than from L2-to-L1 or (b) language-selective access in an integrated lexicon when reading in L2 but not when reading in L1. Neither of these scenarios is compatible with empirical data to date (for reviews, see Brysbaert & Duyck, 2010; Dijkstra & van Heuven, 2002; Grainger et al. 2010; Kroll & Dijkstra, 2002; Kroll et al., 2010). Furthermore, a fully disconnected implementation of Costa et al.'s model also implies either (c) two separate lexica, or (d) one integrated lexicon with entirely language-selective lexical access. However, neither of these options can account for the classic cross-language orthographic/phonological neighbourhood effect (see, e.g. Dirix et al., 2017). Although L2 words with several orthographic neighbours in L1 may end up linked in L2 through learning, under the fully-disconnected model they should not compete for selection because their word forms do not overlap in L2. Hence, although it is reasonable to assume that L2 words may become linked during L2 acquisition because of links between representations existing in L1, partial or full disconnection of L1 would introduce new inconsistencies with a wider array of empirical data.

[3] In fact, Costa et al.'s model includes a 'learning threshold' parameter, Θ_L, to reduce its learning of weak associations. Increasing this threshold, *ceteris paribus*, produces a model that learns to associate translations without learning within-L2 connections.

3.3. Omission of semantic representations

Returning to the original human data, recall that it was generated in the context of a semantic judgement task. Costa et al.'s model, however, omits a fundamental aspect of language which not only seems highly relevant to the behavioural task but which would likely elicit very different results: Semantics. For example, the L2 word 'ham' would become validly associated with words like 'cheese' and 'toast' as a learner of English reads sentences like 'he likes ham and cheese on toast". It is thus likely that –in the course of learning English– words like 'ham' will become associated with words like 'toast' and 'cheese' more than with 'train' because of the lack of semantic association between the former and the latter. Indeed, semantically unrelated English words that overlap in their corresponding Chinese translations would hardly ever see the link between them reinforced during learning as regards meaning acquisition or use, and the effect produced should therefore be much weaker than that of valid and useful within-L2 associations (just like the barely noticeable effect in our new L1-fully-disabled simulation). In our empirical investigations, however, the form overlap effect was about half the size of the semantic relatedness effect (e.g., ~1 vs ~2 μV in Thierry and Wu, 2007) rather than a small fraction of it.

In other words, even though a computational model can be built to simulate connections between pools of neurons on the exclusive basis of form relationship across languages or lexical representations, such a model bears no obvious relationship to language learning in real-life circumstances without taking into account semantic links between words. Thus, it is difficult to imagine how the connection between 'train' and 'ham', once disconnected from L1, would remain as strong as that between words such as 'wood' and 'carpenter', which also have a form overlap through L1 (*mu tou – mu jiang*) but are strongly related semantically in both languages. It thus seems highly likely that word meaning interacts with form overlap between words during learning, a finding that stands in contrast with the lack of interaction observed empirically (see in particular, Thierry and Wu, 2007). This point is particularly relevant considering the fact that most of the experiments conducted by Wu and Thierry involved semantic tasks (i.e., relatedness judgement). A model solely aimed at simulating form overlap effects is thus, at best, incomplete, and seems inappropriate as an account of empirical data obtained in an experimental testing context centred on semantic processing.

Conclusion

In sum, although Costa et al. (2016) propose an intriguing account of Thierry and Wu's (2004, 2007) L1 effects in late bilinguals' L2 comprehension, we find their explanation untenable for the following reasons: (i) A model of semantic judgements that lacks semantic representations is, at best, incomplete; (ii) The mere fact that Hebbian learning *could* lead to acquiring certain associations does not imply that it *must;* (iii) Testing the model after completely –as opposed to partially– severing the links to and from L1 virtually *abolishes any effect of form overlap in L1*; and, finally, (v) The fact that unlearning would eliminate the hypothesised within-L2 connections, unless they were regularly reinforced via intact L1↔ L2 connections, implies that disconnection cannot be the normal state.

Acknowledgements

The authors wish to thank Aina Casaponsa and Manon Jones for comments on previous version of this manuscript.

References

Brysbaert, M. (1998). Word recognition in bilinguals: Evidence against the existence of two separate lexicons. *Psychologica Belgica*. 38 163–175.

Brysbaert M., Duyck W. (2010). Is it time to leave behind the Revised Hierarchical Model of bilingual language processing after fifteen years of service? *Biling. Lang. Cogn.* 13 359–371 10.1017/S1366728909990344

Costa, A., Pannunzi, M., Deco, G., & Pickering, M. J. (2016). Do Bilinguals Automatically Activate Their Native Language When They Are Not Using It? *Cognitive Science*, 1–16. 10.1111/cogs.12434

De Groot, A. M. B., Nas, G. L. J. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *J Mem. Lang.* 30 90-123.

Dijkstra T., van Heuven W. J. B. (2002). The architecture of the bilingual word recognition system: from identification to decision. *Biling. Lang. Cogn.* 5 175–197 10.1017/s1366728902003012

Dirix, N., Cop, U., Drieghe, D., & Duyck, W. (2017). Cross-lingual neighborhood effects in generalized lexical decision and natural reading. *J Exp. Psychol. Learn. Mem. Cogn.*, *43*(6), 887.

Duyck, W. (2005). Translation and associative priming with cross-lingual pseudohomophones: evidence for nonselective phonological activation in bilinguals. *J Exp. Psychol. Learn. Mem. Cogn.* 31 1340–1359.

Duyck, W., Assche, E. V., Drieghe, D., Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: evidence for nonselective lexical access *J Exp. Psychol. Learn. Mem. Cogn.*, 33 663–679.

Grainger J., Midgley K., Holcomb P. J. (2010). "Re-thinking the bilingual interactive-activation model from a developmental perspective (BIA-d)," in *Language Acquisition Across Linguistic and Cognitive Systems*, eds Kail M., Hickmann M., editors. (New York, NY: John Benjamins), 267–284 10.1075/lald.52.18gra

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci. USA. 79*, 2554–2558.

Kroll J. F., Dijkstra A. F. (2002). "The bilingual lexicon," in *Handbook of Applied Linguistics*, ed. Kaplan R. B., editor. (Oxford: Oxford University Press), 301–321.

Kroll J. F., van Hell J. G., Tokowicz N., Green D. W. (2010). The revised hierarchical model: a critical review and assessment. *Biling (Camb Engl)* 13 373–381 10.1017/S136672891000009X

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: a model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227–252. 10.1016/j.cognition.2009.09.007

Paradis, M. (2004). *A Neurolinguistic Theory of Bilingualism*. Amsterdam/Philadelphia. John Benjamins Publishing Company.

Popper K. ([1979] 2009). *The Two Fundamental Problems of the Theory of Knowledge*, Troels Eggers Hansen ed., Andreas Pickel, trans., p. 485.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variation in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Vol. 1. Foundations*. Cambridge, MA: MIT Press

Spalek K., Hoshino, N., Wu Y. J., Damian M., Thierry G. (2014) *Speaking two languages at once*: *unconscious native word form access in second language production*. *Cognition* 133, 226–231. doi:10.1016/j.cognition.2014.06.016

Thierry G. & Wu, Y. J. (2004). Electrophysiological Evidence for Language Interference in Late Bilinguals. *Neuroreport*. 15: 1555-1558.

Thierry, G. & Wu, Y. J. (2007) Brain potentials reveal unconscious translation during foreign language comprehension, *Proc. Nat. Acad. Sci. USA, 104:12530-5.*

van Hell, J. G., Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychon Bull Rev*. 8 780–789.

Wu YJ, Thierry G. (2010a) Chinese-English bilinguals reading English hear Chinese. *J Neurosci.* 30:7646-51.

Wu, Y. J, Thierry G. (2010b) Investigating bilingual processing: the neglected role of language processing contexts. *Frontiers Psychol.* 1:178.

Wu Y. J., Thierry G. (2011) Event-related brain potential investigation of preparation for speech production in late bilinguals. *Front Psychol.* 2: 114.

Wu, Y. J., Thierry G. (2012a) Unconscious translation during incidental foreign language processing. *Neuroimage*. 59: 3468-73.

Wu Y. J., Thierry G. (2012b) How reading in a second language protects your heart. *J Neurosci.* 32: 6485-6489.

Wu, Y. J., Cristino, F., Leek, C., Thierry, G. (2013) Non-selective lexical access in bilinguals is spontaneous and independent of input monitoring: Evidence from eye tracking. *Cognition.* 129(2), 418-425.