Do bilinguals automatically activate their native language when they are not using it?

Do bilinguals automatically activate their native language when they are not using it?

Albert Costa[1,2], Mario Pannunzi[2], Gustavo Deco[1,2] & Martin J. Pickering[3]

1. ICREA, Barcelona, Spain

2. Department of Information and Communication Technologies, Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain

3. Department of Psychology, University of Edinburgh, UK

Abstract

Most models of lexical access assume that bilingual speakers activate their two languages even when they are in a context in which only one language is used. A critical piece of evidence use to support this notion is the observation that a given word automatically activates its translation equivalent in the other language. Here, we argue that these findings are compatible with a different account, in which bilinguals "carry over" the structure of their native language to the non-native language during learning, and where there is no activation of translation equivalents. To demonstrate this, we describe a model in which language learning involves mapping native-language phonological relationships to the non-native language, and show how it can explain the results attributed to automatic activation of translation equivalents.

Bilinguals sometimes make use of both of their languages at the same time, for example when they change language mid-utterance (i.e., code-switch) or change language when answering a question. In such (two-language) contexts, they may well activate their two languages in parallel. But most of the time, bilinguals make use of just one language, for example when writing, listening to the radio, or having many conversations. In such (one-language) contexts, we might naturally assume that they activate the relevant (target) language but not the irrelevant language.

This natural assumption has been challenged in recent years. In particular, the current dominant view is that bilingual language processing involves rapid and extensive interaction between languages even when one language is wholly irrelevant, that is in one-language contexts. In effect, according to this view, bilinguals constantly activate the word being processed in the target language and also the corresponding translation in their other language, with activation passing between translation equivalents (e.g., Thierry & Wu, 2007). That is, the presentation of the word "table" will lead to the activation of its translation equivalent "mesa" ("table in Spanish). Here, we propose that the experimental data used to support such language co-activation are in fact compatible with an alternative account based on learning, in which activation does not need to pass between translation-equivalent words. In this paper, we describe this account, present a computational model of our proposal, and discuss its implications.

Language processing in bilinguals: On-line interaction or the remnants of learning?

How can we tell if a bilingual activates the irrelevant language in a one-language context? Obviously, we cannot conduct an experiment that involves both languages (e.g., cross-linguistic picture-word interference or language switching). Similarly, occasional cross-linguistic intrusions need not be informative about regular processes of activation (see Costa, La Heij, & Navarette, 2006, for a discussion of the difficulties about finding the appropriate experimental contexts to test this hypothesis).

Another approach is therefore to consider processing within the target language only. For example, some words in the target language and non-target language are cognates (related in form and meaning, e.g., *guitar-guitarra* in Spanish). There is good evidence that bilinguals in a one-language context process cognates differently from non-cognates, and differently from the way that monolinguals process such words. Specifically, studies show a cognate advantage under some conditions at least (Costa, Caramazza, & Sebastian-Galles, 2000). This finding indicates some form of link between the two languages. Most theories (Costa, Miozzo, & Caramazza, 1999; La Heij, 2005; Green, 1986; 1998) assume on-line activation of the non-target language. For example, Costa et al. (2000) proposed that Spanish-English bilinguals name a picture as *guitar* quickly because they activate the phonology from both languages, and the activation of *guitarra* facilitates its cognate *guitar*.

But another possibility is that bilinguals learn different representations for cognates and non-cognates. As cognates are more similar to their translations than non-cognates are, they are likely to be easier to learn, and may then be represented more prominently than non-cognates. Specifically, cognates would be more accessible than non-cognates, after controlling for other characteristics that affect response time (e.g., phonological properties, word frequency, and age of acquisition). If this alternative

explanation is correct, then the comparative ease of processing cognates would not reflect on-line activation of the non-target language but would instead be the result of learning. More generally, within-language processing effects would not, by themselves, demonstrate on-line interaction between languages.

Perhaps the most compelling way to show parallel activation of the two languages is to demonstrate that in monolingual contexts, processing a word in one language ("table") activates the corresponding translation word in the other language ("mesa"). We will argue, however, that the current evidence on this issue is also compatible with a learning-based account.

Do bilinguals using one language activate the corresponding translation equivalents?

The goal of the present article is not to review all the experimental data that has been used to support the presence of language co-activation in bilinguals. Rather, we focus on some of the best-known evidence suggesting such co-activation, and use it as an example of how a different account that dispenses with such co-activation could explain the experimental observations. In our view, the key evidence relates to the activation of translation equivalents.

In an ingenious study, Thierry and Wu (2007) asked Chinese-English bilinguals to judge whether two sequentially presented words in their second language, English (e.g., *train* and *ham*) had related meanings. In some cases, the pairs were semantically related and in others unrelated. The crucial manipulation, however, was the formal relationship that the translation equivalents of these words had in the first language of the participants, Chinese. In both a spoken and a written experiment, they found reduced N400 amplitude when these words had a form-related translation (in this case,

huo che and *huo tui*) in the first language, Chinese, versus when they did not. This effect was observed irrespective of whether the word pairs were semantically related or not. No such effect occurred when monolingual English participants encountered the same English words, but a similar effect did occur when monolingual Chinese participants encountered the translations in Chinese. The authors argued that the presentation of given word in the second language of the participants automatically activates its translation equivalent in their first language. The similarity between the word forms of the two translation equivalents drives the reduction in the N400 component. Note that the title of the paper assumes on-line translation (*Brain potentials reveal unconscious translation during foreign-language comprehension*), but we are not concerned with whether the mechanism should be interpreted in this way or simply in terms of co-activation. Instead we focus on their conclusion: "In sum,…results reveal an automatic translation processing (….). This finding provides an account for parallel, language nonselective activation models of bilingual word recognition" (p. 12534).

We argue that these Thierry and Wu's (2007) results can be explained without activation of translation equivalents, if we assume that the English lexicon of the Chinese-English bilinguals is fundamentally different from the English lexicon of native English speakers. Specifically, the lexical organization of a second language may carry remnants of the way the first language is structured. Informally, the Chinese-English bilinguals initially represent *huo che* and *huo tui* as having a form-based association (i.e., *huo*). To learn the English *train* and *ham*, they "copy" their representations from their Chinese lexicon to their developing English lexicon, including the association between their forms. As a consequence, the corresponding English translations also become associated. In other words, their English lexicon encodes traces of the

relationship between the translation words in their native language, Chinese. On this view, the effects that Thierry and Wu (2007) assumed are caused by the activation of translation equivalents could actually be due to relationships *within* their (non-native) English lexicon. Their effects would reveal the structure of their English lexicon rather than parallel activation of English and Chinese. Thus the critical issue here is how two sets of representations associated with the native and non-native language interact in the course of learning a second language. Before going into the details of this account, let us illustrate with two very different examples, how learning a new set of representations may alter other sets of representations that have been already established.

The first example comes from an analogous, and unresolved, controversy about the way in which orthography influences spoken-language processing. In an auditory lexical-decision task, Ziegler and Ferrand (1998) found that people had more difficulty with words whose rimes could be spelled different ways (e.g., *beak*) than with words whose rimes could be spelled only one way (e.g., *luck*). This finding could reflect on-line effects of orthography on phonology. If so, orthographic codes would be automatically activated during speech comprehension. This account assumes parallel activation of different codes, in this case orthography and phonology rather than a native and a non-native language. But the finding may instead be due to "phonological restructuring," by which orthography "contaminates phonology during the process of learning to read and write, thus altering the very nature of the phonological representations themselves" (Petrova, Gaskell, & Ferrand, 2011, p. 2). On this account, the acquisition and development of orthographic representations (i.e., literacy) changes the nature of the already established phonological representations through learning, and orthographic-phonologic interactions would not occur during speech processing. In

sum, effects of orthography during spoken-language processing might be due to on-line interaction, but they also might be due to remnants of learning.

The second example comes from the study by Warker and Dell (2006) in which learning a new phonological "mini-grammar" was affected by the already existent phonological system of the participants' native language, English. Indeed, the authors simulated the learning of this new grammar by "copying" the phonological structure of English and then training such structure with the new mini-grammar. This resulted in transfer from the already learned system (English) to the new system that was being learned.

How do unrelated representations end up being related as a consequence of learning?

A fundamental step in second language word learning is that of linking new phonological forms to conceptual information. This will allow the fundamental purpose of learning a new language, namely the ability to convey meaning. As a consequence, it is reasonable to assume that the new acquired words will be structured according to semantic relationship, resembling the organization of the first language. However, the question is whether during second language learning, there may be other factors that affect the organization of the lexicon. We think that one of those factors can be the way lexical items in the lexicon of the first language are already organized. In this way, lexical items that are in principle not semantically related may end up being related by virtue of inheriting the relationship that their translation counterparts have at the formal level.

Our alternative explanation of the results of Thierry and Wu (2007) without appealing to parallel activation hinges in the assumption that during learning a second

language, there is transfer between the structures of the two lexicons. This means that unrelated words in one language (e.g., *train* and *ham*) may end up being related, by virtue of the relationship that the corresponding translations have in the other language (e.g., *huo che* and *huo tui*). But how does this transfer come about? As we will see, the mechanisms that account for this transfer are similar to those embraced by researchers assuming the parallel activation of the two languages. The first mechanism is that of spreading activation between related representations, whether the relationship is semantic or formal. Second, representations that are activated at a similar point in time end up developing connections between them.

According to the mechanism of spreading activation, the Chinese learner of English who was learning the word *Train* would first activate its translation equivalent (*huo che*) by virtue of their semantic relationship. The activation of *huo che* would then spread to *huo tui*, by virtue of their phonological relationship. *Huo tui* would in turn activate its translation equivalent, *ham*, by virtue of their semantic relationship. The assumption of spreading activation and its implications are widely accepted by many models of language processing (e.g., Levelt, Roelofs, & Meyer, 1999; Dell, 1986). Namely, activation spreads between representations that are related in some way or another. Indeed, spreading activation, at least, between semantically related representations is also assumed by Thierry and Wu (2007). According to this principle, at some point during the learning process it follows that the presentation of the word *train* will lead to the activation of the unrelated word *ham*.

The second mechanism posits that representations that fire together end up forming connections. This is the basis for Hebbian models and is also assumed in classical connectionist models (Dell, 1986; Fusi, Annunziato, Badoni, Salamon, & Amit, 2000; Hebb, 1949). If so, for two pools of neurons to develop a connection they

need to become activated at a similar time during processing. As we have seen, spreading activation predicts that *train* and *ham* should be activated in close temporal proximity, and consequently it is reasonable to assume that they should develop a connection.

This provides a simple way in which two phonologically and semantically unrelated words in a non-native language (*train* and *ham*) can be end up being connected through transfer from the corresponding translation words in the native language (*huo tui* and *huo che*). In this way, we can explain how the structure of one language can be copied on the structure of the other.

Refining the conditions leading to effects of translation equivalents

Our account can also explain other studies that have been interpreted in terms of the activation of translation equivalents. Wu and Thierry (2010) distinguished sound and spelling repetition and found effects similar to Thierry and Wu (2007) when the translations were related in sound but not spelling. They argued that processing a second language activates the sound, but not the spelling, of native language translations (with the title of the paper being *Chinese-English bilinguals reading English hear Chinese*). This is a rather interesting finding that puts constraints on the types of codes that are linked across languages. Another factor that seems to affect translation equivalents is emotional valence. Wu and Thierry (2012a) observed a reduced N400 amplitude when the English prime word had positive or neutral affective valence, but no effect when it had negative valence (e.g., *failure*). These two modulating factors of cross-language activation are worth exploring further, but at present do not posit a challenge for our learning-based account. Specifically, our account proposes that the process underlying learning is based on phonology rather than orthography

(presumably because phonology is more basic to lexical representation), and in addition a negative valence may interfere with learning. In any case, a learning-based explanation can account for the different effects of phonology and orthography, and the modulatory effects of valence, just as straightforwardly as an activation-based explanation.

In a rather different study, Wu, Cristino, Leek, and Thierry (2013) had Chinese-English bilinguals search for strings of circles (or squares) in a grid that also contained three English words. Participants looked at English words more often if their Chinese translation phonologically resembled the Chinese word for circle (or square) than otherwise (see also Wu & Thierry, 2012b). They suggested that incidental word processing leads to activation of the non-target language. This is completely consistent with the learning account put forward here, since the effects are assumed to reveal the structure of the lexicon despite any intentional or incidental processing. That is, the effects arise because of the way the first language has shaped the organization of the second language lexicon, and consequently these effects should be independent, to some extent, of attentional factors.

Moreover, other groups of researchers have made similar claims to Thierry, Wu, and their colleagues. Thus, Zhang, van Heuven, and Conklin (2011) also argued for "fast automatic translation" in a study where Chinese-English bilinguals made lexical decisions to English words that were preceded by a briefly presented masked English prime word. Responses were faster when the Chinese translations of the prime and target words had the same first character than when they did not. (There was no effect

of second-character repetition.) The authors argued that participants conducted fast and automatic translation of the English words into Chinese. Finally, Morford, Wilkinson, Villwock, Pinar, and Kroll (2011) had ASL-English bilinguals judge whether English word pairs were semantically related, and found faster responses when the ASL translations of those words were related in form than when they were not; these effects did not occur with participants who were not bilingual in ASL. In all these studies, the researchers interpreted the experimental observations as revealing the co-activation of the two languages of a bilingual. But again the effects can be explained in terms of learning.

Before describing the computational model of the alternative learning-based explanation of these phenomena, let us reiterate the idea behind it. In our view, given the likely interaction between second language and first language representations during learning, the organization of the lexicon of the second language may retain traces of the first language (see also Zhao & Li, 2010). If so, the organization of a speaker's second language (say English) would depend on the properties of the speaker's native language (say Chinese or Spanish), and, in particular, would be different from that of a monolingual English speaker. Specifically, lexical items that do not appear to be related in the second language would develop related representations by virtue of the relationship (phonological in this case) of their corresponding translations in the native language.

Description of the computational model

The main goal of our model is to show how the effects interpreted as revealing activation of translation equivalents could be due to the influence of the native language

(L1) on the structuring of the non-native language (L2) during learning. Our model makes two main assumptions: 1) parallel activation of the two languages occurs during second language learning, and 2) activation becomes restricted to one language when a given proficiency level in the second language is attained. The first assumption is quite standard in models of bilingual language acquisition. With respect to the second, the notion that bilingual proficiency may affect the type of processes involved in bilingual lexical access is often entertained (e.g., Costa & Santesteban, 2004). Once the model has been trained simulating learning the L2, we can assess the activation of supposedly unrelated representations in the L2 while removing any on-line influences between L1 and L2. This allows the assessment of whether the resulting structure of the L2 is influenced by the already existing L1 structure. More critically, it will indicate whether activation within the resulting L2 lexical network can reproduce the effects observed in Thierry and Wu (2007) without parallel activation of the two languages. In more lay terms, it is as if we remove any online influence of the L1 during L2 processing.

Note that this way of testing the model seeks to demonstrate a proof of concept that it is possible explain the results of Thierry and Wu (2007) without assuming parallel activation of the two languages. In other words, we would show that removing any possible co-activation of the two languages may lead to the same results as allowing full co-activation. This does not mean that the abrupt removal of language co-activation reflects the real processes undergone by bilinguals. It is very likely that activation is more gradually restricted, in a way that likely reflects increasing proficiency and automaticity. However, we decided to test the model in the strictest and more demanding conditions, and this is why we abruptly removed any contribution of the non-active language.

Abstract description of the main assumptions of the model

The model included six words whose activity was each simulated by an independent pool of neurons. Three words corresponded to the speaker's L2 (English) and were unrelated in meaning and form (*train, ham, apple*). The other three words were their translations in L1 (Chinese) (*huo che, huo tui, pin guo*). Crucially, although all three Chinese words are unrelated in meaning, two of them are related in phonological (or orthographic) form.

The different pools representing words are connected through excitatory plastic connections that simulate (to some extent) synaptic linkage. At the beginning of learning, the strength between connections that do not hold a linguistic relationship was set to zero. Hence, at the beginning of learning the L2 the only actual functional connection was that between two words in the L1 (*huo che* and *huo tui*). All other connections were set to 0. We are concerned with how the strength of all these connections varies as a consequence of L2 learning. However, the critical issue is the strength of those connections between apparently unrelated words in the second language (here, between *train* and *ham* and between *train* and *apple*).

The resting state of all the pools was the same when no linguistic activity was simulated. Linguistic activity during learning was simulated as follows. When a word was encountered for the first time in the second language (e.g., through the presentation of a picture or a word), an excitatory current was injected into the pool of neurons corresponding to that word. Furthermore, a current was also injected into the corresponding translation-equivalent word in the first language, thereby simulating parallel activation of the two languages during learning. Then, activation spread to representations whose connections were higher than 0. In this context, the first time the model encountered the L2 word *train*, there was activation of (the representation of) its translation word in the first language (*huo che*), and then the activation of *huo che*

spread to the form-related word *huo tui*. Importantly, the level of activation of the presented word was always higher than the activation of related words. This served to implement the fact that this word was presented rather than activated as a result of its connections with a different word.

Now the question is how this parallel activation of the two languages affects the development of the strength of the connections across learning. The model assumes that the strength of the connections between representations develops as a result of a Hebbian learning mechanism. That is, the link between the representations of *huo che* and its translation *train* increases in strength as the Chinese speaker learns English because both words have the same meaning. Increasing the strength of the connections results in spreading activation, where the activity of one lexical representation spreads to other lexical representations which then become activated.

This learning then leads to an increase of the strength of the connections between translation words, and also between seemingly unrelated words. For example, presentation of *huo che* leads to the activation of both *train* and *huo tui*. In turn, *huo tui* spreads some activation to its translation *ham*. This means that the representations of *ham* and *train* are co-activated during learning, and this co-activation due to Hebbian learning increases the connection between them. Hence, the functional connectivity between representations of words that are neither semantically nor phonologically related (*train* and *ham*) increases. This increased connection of course does not develop in an English monolingual, since there is no co-activation of *ham* and *train*.

We simulate three aspects of Thierry and Wu's (2007) study. First, we consider when learning has not occurred yet. This initial state also corresponds to a monolingual speaker, as the strength of the connections between languages is zero (and therefore can be compared to Thierry and Wu's monolingual condition). This context serves as a

control against which to compare the results when the model has been trained. Second, after learning, we simulate their study of bilinguals under conditions in which the languages can be activated in parallel. These conditions are of course compatible with Thierry and Wu's account. To do so, we compare the activity of the pool of neurons corresponding *ham and apple* when *train* is presented. Third, we simulate their study in a modified model that prevents parallel activation of the two languages after learning, so that there is no L1 activation when the L2 is presented. We do so by "switching off" the on-line connections between the languages (the arrows that connect the boxes in the lower panel of Fig. 1). In cognitive terms, switching off these connections means that there is no on-line translation, and that processing occurs only in the target language, in this case the L2. Our critical concern is with the activity of the pools of neurons corresponding to *ham and apple* following presentation of *train* in this final situation: Does activation transfer between *ham* and *apple* without on-line activation of their translations? As already discussed, this is an abrupt disconnection that does not necessarily reflect a natural situation in bilinguals. Probably, the disconnection (or lack of parallel activation) is a more gradual process that is likely the result of increasing proficiency and automaticity. We decided to introduce such an abrupt disconnection in our model as a proof of concept.

Activities during learning

The presentation of a target L2 word was stimulated by an external current (average ν_H, standard deviation $\Delta\nu$) to the corresponding pool of neurons. Then, we also applied an external current to the pool of neurons corresponding to its translation word

in L1 (ν_{H2}, standard deviation $\Delta\nu$). All the other pools of neurons received background stimulation (average ν_{VL}, standard deviation $\Delta\nu$).

With these assumptions and after some training, we expect the following activity pattern when the target word *train* is presented. First, the activity of the target's translation *huo che* should be relatively high. Second, the activity of a word that is phonologically related to this translation (*huo tui*) should be relatively high too. Third, and crucially, the activity of the translation of *huo tui* (*ham*) should have an intermediate value. Finally, the activity of words unrelated in both meaning and form to the target word and its translation should be relatively low (*apple, pin guo*).

Note that we hypothesized that all the pools encoding these words reach different levels of activity, and we set the injected current to *train* higher than that to its translation *huo che*. We assume that the activity of each pool is a linear function of all external inputs – that is, the other words' pools of activity and the external stimulus (word presentation). The connection strengths between pools were set all to zero, apart for the connection between the words *huo tui* and *huo che*, which was set to the higher value c_{Ph} because of their phonological similarity. The recurrent connections were fixed to zero for simplicity. The values of all the parameters used are reported in Table 1.

Hebbian learning

Learning, based on a simplified form of Hebbian reinforcement, takes place at the level of the connectivities between pools: When two pools have high activities, their connectivity increases. The assumptions made for the training are: 1) Each L2 word is pronounced various times, and therefore its activity is high; 2) Every time a word is pronounced, the connectivities between it and all the other words change as a function

of both its activation and their activation; 3) The speed of learning (rate of connection increase) is a sigmoidal (s-shaped) monotonic function of the activity of the L2 word; 4) Learning takes place only when a L2 word's activation is above a given threshold; 5) Learning is probabilistic and its strength decreases over time, with strength being $2/(1+\exp(n/N_{max}))$, where n is the number of times the word is pronounced (or seen), and N_{max} is the parameter governing the decreasing speed.

We set the probability to increase the connection between two pools to be proportional to the normal cumulative distribution function. This choice was made for simplicity and in order to make learning follow a sigmoidal monotonic function. The function has therefore two parameters that are the mean (Θ_L) and the standard deviation (β_L) of the normal distribution function, together with a proportionality parameter α_L.

Results

The model allows us to assess both the activation of pools of neurons across time and the strength of the connections between different pools of neurons (see Fig. 2). Regarding the strength of the connections between the pools of neurons, panel A of Figure 2 shows that at the beginning of training the only connection whose strength increases significantly is the connection between translation words (*train* and *huo che*). Also, the critical connections between *train* and *ham* and between train and *apple* begin at zero. However, at the end of the learning we can see differences between these latter connections, with the strength being greater between *train* and *ham* than between *train* and *apple*. This means that learning leads to an increase in the connections between two unrelated representations.

As noted above, we simulated Thierry and Wu's (2007) study in three different situations. Recall that the critical comparison is the activation reached by the pools of neurons corresponding to *ham* and *apple* when an unrelated word *train* is presented. Panel B of Figure 2 shows at the beginning of training the distribution of activities of the pools *ham and apple* during the presentation of the target word *train*. We can see that at the beginning there is no difference between the activities of these two pools. We repeated the presentation of *train* 8000 times. During the course of training, *ham* becomes more and more activated. That is, learning *train* ends up activating the unrelated word *ham* to a larger extent than *apple* (panel C).

The critical situation in which to evaluate Thierry and Wu's (2007) proposal is after training and when any influence of the L1 is removed. To model this, we kept the connections' values that resulted from training but now removed all L1 representations. We then activated *train* (8000 pronunciations) and measured the dynamics of the pools of neurons corresponding to the two words *ham* and *apple*. As predicted, activating *train* alters the activity of the seemingly unrelated word *ham* relative to the word *apple* (see panel D of Fig. 2). This modulation of the word *ham* relative to *apple* is not due to parallel on-line activation of the L1, since those representations have been removed from the model. Instead, the modulation is due to the relationships within the representation of the L2. As it can be appreciated when comparing panels C and D of Figure 2, the model's results are very similar, regardless of whether parallel activation of the L1 is present or not.

General Discussion

The main goal of our model was to show how effects that have been interpreted as revealing activation of translation equivalents could instead be due to the influence of the native language on the structuring of the non-native language during learning. To do this, we simulated the results reported by Thierry and Wu (2007) in a model that allows co-activation of the two languages during learning, but that restricts activation to only one language at some point after learning.

Importantly, this restriction in activity occurs only after learning has taken place, with cross-talk between the two languages occurring during learning. This cross-talk causes the lexical organization of the L2 to contain traces of the lexical organization of the L1. On this account, Thierry and Wu's (2007) findings reflect this lexical organization rather than parallel activation (i.e., on-line cross-talk). The same argument applies to other studies that use similar logic to Thierry and Wu (Morford et al., 2011; Wu & Thierry, 2010, 2012; Zhang et al., 2011).

To assess the feasibility of these claims to account for the experimental observations, we constructed a toy model that implemented two main assumptions: 1) a learning process in which there is a parallel activation of the two languages, and 2) a process whereby activation is restricted to one language when proficiency level in the L2 is sufficiently high. After the learning phase, we "turned the model monolingual" by removing the presence of L1 representations.

Note that our model is silent about how this restricted activation comes about when proficiency is attained. It does not demonstrate how this restricted activation is implemented (e.g., via inhibition) and its time-course (which it is unlikely to be as abrupt as implemented here). These are issues to be investigated in future work. What it is important for our purposes is that the model can simulate results supporting activation of translation equivalents, when such translations are removed. As a result of Hebbian

learning, the structure of the L1 representations was (partly) mapped to the L2 representations. This results in a L2 lexical structure that depends to some extent of the lexical structure of the L1. On our account, the structure of the English lexicon is different for the native speaker of English, the native speaker of Chinese, and the native speaker of Spanish. The model was able to reproduce Thierry and Wu's (2007) key finding of within-L2 priming between words that were phonologically unrelated in the L2 but phonologically related in the L1. Therefore, it opens the possibility that their results are not due to the parallel activation of the two languages but rather to the interaction or transfer between the structures of the two lexicons during learning.

So, where do we stand with respect to the presence of co-activation of the two languages? We have not, of course, demonstrated that there is no parallel activation of the two languages in a one-language context, nor that there is no activation of translation equivalents – and indeed this was not our intention. It is possible that such parallel activation does occur (and of course it may also occur in two-language contexts when people switch between languages), and that therefore the interpretation of Thierry and Wu (2007) in terms of activation of translation equivalents is partly correct. What we have just shown is that there is an alternative way to interpret the results that dispenses with language co-activation. Hence, caution needs to be exercised when using such results to support the presence of language co-activation, since the results are consistent with another interpretation. Indeed, teasing apart these two interpretations may prove difficult, since it would be necessary to find the conditions that allow us to test the parallel activation of the two languages without being sensitive to the potential re-structuring of the L2 as a consequence of the L1, and vice-versa. At the moment, Thierry and Wu's results are consistent with the two alternative interpretations.

As discussed in the Introduction, it was not the goal of this article to review all the studies that have explored the issue of parallel activation of the two languages. We have focused on simulate one of the most compelling phenomena that has been repeatedly used to support the idea of activation of translation equivalents during language processing. However, our account could be developed to address other results that have been interpreted in terms of the parallel activation of the two languages (or at least the notion of activation of translation equivalents). Let us consider one of them in detail.

Marian and Spivey (2003) presented Russian-English bilinguals with four objects (a shark, a balloon, a napkin, a horse) and instructed them in English to direct their attention to a target (*pick up the shark*). The Russian name for one of the other objects (*sharik*, meaning balloon) is phonologically similar to the English name for the target (*shark*). Participants tended to look at this distractor picture (balloon) more than to other pictures. According to the authors, this result indicates that people automatically activate the mental lexicons for both languages in parallel. But the effect can also be explained by our proposals, in which parallel activation is present during learning but absent when proficiency increases. The argument is similar to the one developed above: The Russian speaker co-activates *sharik* when learning *shark* because they are phonologically related, and *sharik* in turn activates *balloon*. Once such connection has been established, then the representations of *shark* and *balloon* would tend to be activated together even if online activation is restricted to English. In other words, Russian L1 speakers would treat *shark* and *balloon* as having related English representations and so Marian and Spivey's effects can be explained without on-line activation of Russian. Similarly, we propose that the lexical restructuring might be used to explain other effects that have been used to support on-line parallel activation, such

as the effects of cognates and false friends in reading. Note, however, that our alternative explanation does not imply that bilingual language processing does not involve control processes. The question it raises is at which level this control is exercised and how it interacts with the level of activation of the lexical representations belonging to each language.

We conclude by pointing out potential caveats or limitations of our model. First, the bilingual experience comes in many forms and many variables may affect the cognitive structures that result from learning and using two languages. Processing may be affected by the age at which the two languages are acquired, level of proficiency, regularity of use, or similarity between the two languages. Our model has not considered these variables. But all of these factors can be interpreted in terms of learning just as much as they can be interpreted in terms of on-line activation. Notwithstanding, future research is needed to assess how such variables may modulate this cross-talk during learning (see Zhao & Li, 2010).

Second, we have not addressed how long lasting can be the footprint of the first language on the structure of the second language, once lexical activation is restricted to only one language, as it is assumed here. Arguably, if lexical activation is increasingly restricted to one language, then the links between unrelated representations (via the activation of the other language) would weaken across time (*train* and *ham* would activate each other with less intensity). In other words, the system might unlearn so that footprint of the first language on the second language would reduce over time (at some rate). In addition, it is possible that higher proficiency in a second language would lead to greater autonomy between the two languages, as a result of the reduction in cross-language activation. As a consequence, the influences of L1 on L2 would be less obvious as proficiency increases.

The computational simulation presented here also opens several questions for further research. For example, we could investigate whether the effects are reversible, as might occur when speakers stop using their L1 regularly. In such cases, the connections between L2 lexical representations that are related only by the properties of the corresponding L1 translations are not refreshed regularly, and so may disappear via depotentiation. That is, the L2 structure may dynamically change in such a way that reduces the influence of L1 lexical structure. We predict that the L2 lexical network of this type of bilinguals will be much less affected by the L1 lexical network, and that they may not show the experimental effects often interpreted as automatic translation.

Finally, we also want to mention that in the same way as the L2 representations may carry traces of the organization of the L1, it is also possible L1 representations may carry traces of the organization of the L2. That is, following the interaction between the two languages during learning and as a consequence of Hebbian learning, the acquisition of a new language may alter the structure of the L1. If so, the lexical organization of speaker's first language would depend on whether they know also another language, so that English lexicon of English-Spanish speaker would have a different structure from that of an English-Mandarin speaker or a monolingual English speaker. But such L2-on-L1 effects have not, to our knowledge, been demonstrated so far. In conclusion, we have shown that evidence for on-line activation translation during one-language processing is also compatible with a learning account in which no on-line activation occurs.

References

Costa, A., La Heij, WL., Navarrete E. (2006). The dynamics of bilingual lexical access. *Bilingualism: Language and Cognition, 9*, 137-151.

Costa A, Caramazza A, & Sebastian-Galles N. (2000). The cognate facilitation effect: implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(5), 1283-96.*

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection? *Journal of Memory and Language, 41, 365-397*.

Dell, G.S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283–321.

Fusi, S, Annunziato, M., Badoni, D., Salamon, A., & Amit, D.J. (2000). Spike-driven synaptic plasticity: theory, simulation, VLSI implementation. *Neural Computation,* 12(10), 2227-2258.

Green, D.W. (1986). Control, activation and resource: a framework and a model for the control of speech in bilinguals. *Brain and Language, 27*, 210-223.

Green, D.W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition, 1*, 67-81.

Hebb, D.O. (1949). The Organization of Behavior. New York: Wiley & Sons

La Heij , W. *(*2005*)*. Lexical selection in monolinguals and bilinguals. In J. F. Kroll A. M. B. de Groot (Eds.), Handbook of bilingualism (pp. 289-307). New York:

Levelt, W.J.M., Roelofs, A., & Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–38.

Morford, J. P., Wilkinson, E., Villwock, A., Piñar, P., & Kroll, J. F. (2011). When deaf signers read English: Do written words activate their sign translations? *Cognition*, *118*(2), 286-292.

Thierry, G., & Wu, Y.J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences, 104*(30), 12530-12535

Petrova, A., Gaskell, M.G., & Ferrand, L. (2011). Orthographic consistency and word-frequency effects in auditory word recognition: New evidence from lexical decision and rime detection. *Frontiers in Psychology, 2,* 1-11.

Warker, J.A., & Dell, G.S. (2006) Speech errors reflect newly learned phonotactic constraints. Journal *of Experimental Psychology: Learning, Memory, and Cognition, 32* (2); 387-398

Wu, Y. J., & Thierry, G. (2010). Chinese–English bilinguals reading English hear Chinese. *The Journal of Neuroscience*, *30*(22), 7646-7651.

Wu, Y. J., & Thierry, G. (2012a). How reading in a second language protects your heart. *The Journal of Neuroscience*, *32*(19), 6485-6489.

Wu, Y. J., & Thierry, G. (2012b). Unconscious translation during incidental foreign language processing. *NeuroImage*, *59*(4), 3468-3473.

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.

Zhang, T., van Heuven, W. J., & Conklin, K. (2011). Fast automatic translation and morphological decomposition in Chinese-English bilinguals. *Psychological Science*, *22*(10), 1237-1242.

Zhao, X., & Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. *International Journal of Bilingual Education and Bilingualism, 13*, 505–524.

Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory recognition. *Psychonomic Bulletin & Review, 5*, 683-689.

Table 1

ν_H	40 au
ν_{H2}	15 au
ν_{VL}	4 au
$\Delta\nu$	2 au
Θ_L	25
N_{Max}	6000
β_L	6 au
α_L	0.008
c_{Ph}	0.60

Figure 1. Schematic representation of the L1 and L2 words and their connections at the beginning (top) and at the end (bottom) of the learning. Each rectangular box represents a language (Chinese and English). The connections between the pools of neurons corresponding to the words in the two languages are the result of L2 learning (solid black lines in the bottom). The dashed black arrows linking *huo che* and *huo tui* represent the enhanced connections that are due to their phonological relationship and are equally represented at the end and at the beginning of the learning. The thick black arrows linking *train* and *ham* represent the enhanced connections that develop as a result of their translations' phonological relationship (see text). The gray arrows represent the link between those pools of neurons lacking of enhanced connections.

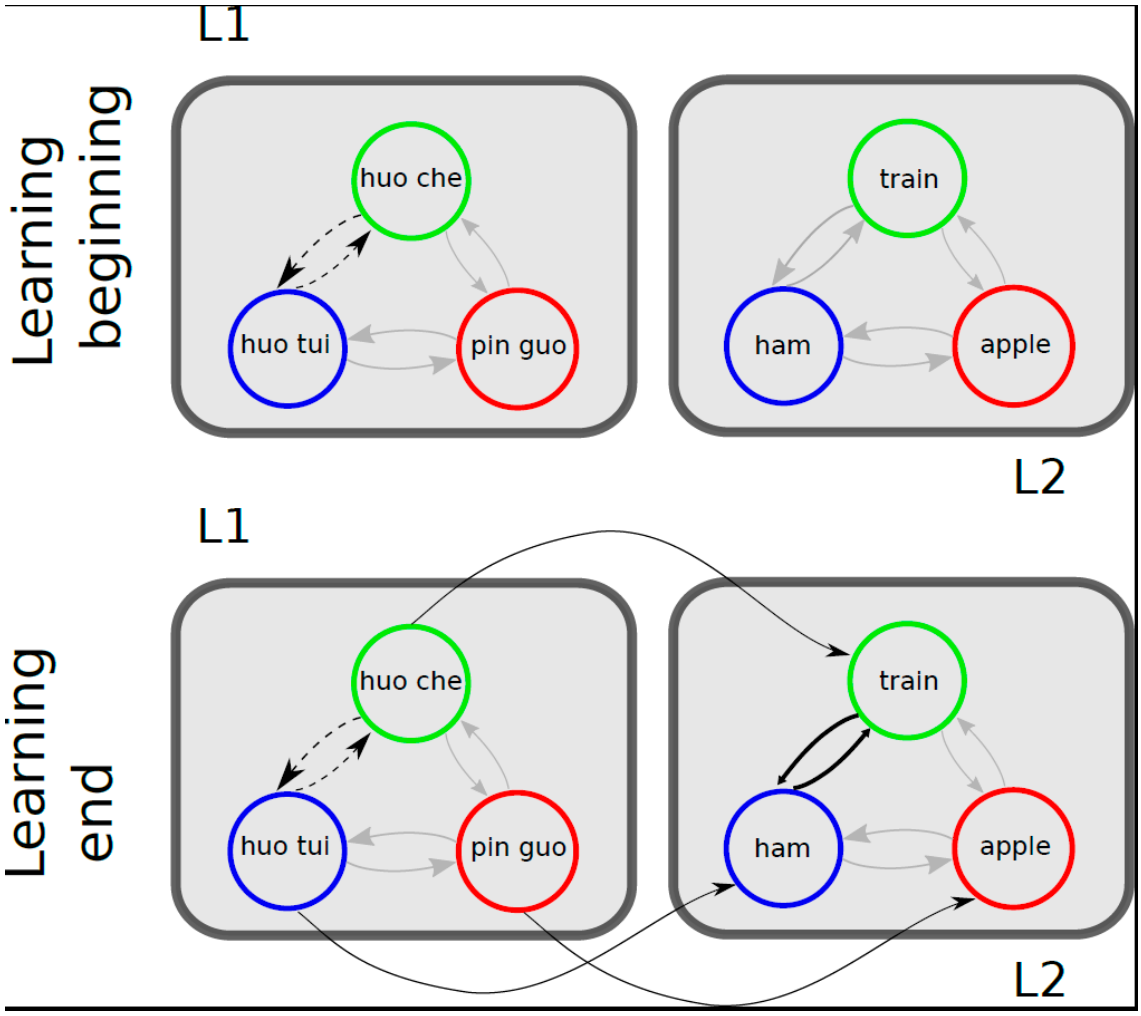
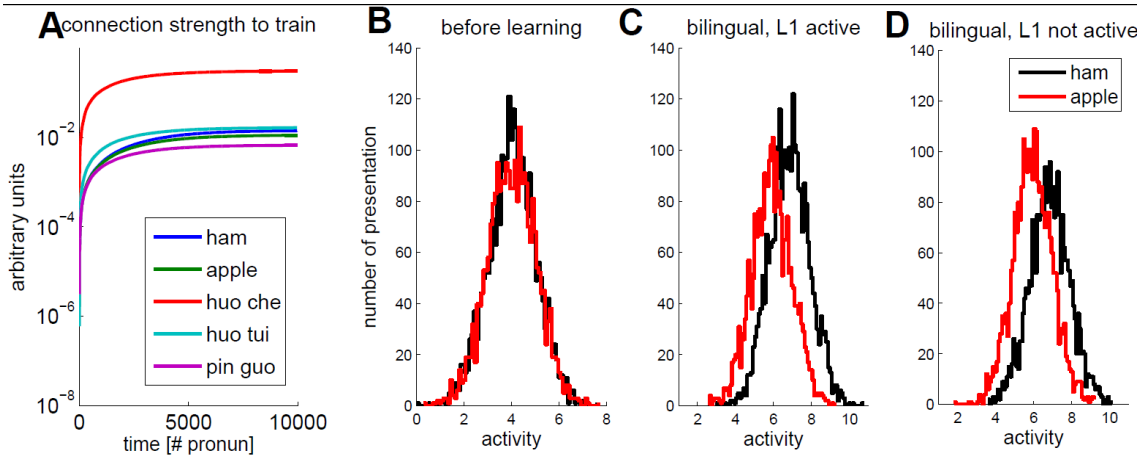


Figure 2. Connectivities and activities evolution throughout the learning. The panel A shows the evolution of the connections to the word *train* from the other five pools during the learning period. Panels B-D represent the activity distributions for the words' pool *ham* and *apple* following presentation of the word *train* in three different situations: before learning (panel B), after learning (panel C), and during the test of Thierry and Wu's (2007) critical experiment (panel D). After learning (Panel D), the word *train* is presented alone (without its L1 translation *huo che*), and yet the word *ham* has a higher activity level than *apple*. We used arbitrary units (a.u.) for the y-axis, and the unit of measurement for the time is the number of times the L2 word was presented (pronounced or seen).



Acknowledgements:

This research was approved by the ethics committee of the Spanish Ministry of Economy and Finance, which funded this study. (PSI2014-52181- P), and Consolider INGENIO CSD2007-00012) awarded by the Spanish Government; by one grant from the Catalan Government (SGR 2014-1210); and by one grant from the European Research Council under the European Community's Seventh Framework (FP7/2007-2013 Cooperation grant agreement 613465-AThEME). AC is supported by the ICREA institution and the Center for Brain and Cognition. MP and GD were supported by ERC Advanced Grant DYSTRUCTURE (n. 295129), theFP7- FET-Flagship Human Brain Project (n 604102), Plan Estatal de Fomento de la investigación Científica y Técnica de Excelencia (PSI2013-42091-P), Grup de Recerca Cognitve and Computational Neuroscience (CCN) 2014SGR856 (AGAUR), MJP acknowledges the support of an grant from the Royal Society of Edinburgh (RSE-NSFC Joint Project Scheme).