Bilingualism: Language and Cognition 9 (2), 2006, 137–151 © 2006 Cambridge University Press doi:10.1017/S1366728906002495

The dynamics of bilingual lexical access*

ALBERT COSTA
GRNC, Parc Científic Universitat de Barcelona & Hospital Sant Joan de Déu
Dept. de Psicologia Bàsica, Universitat de Barcelona
WIDO LA HEIJ
Leiden University, Cognitive Psychology Unit
EDUARDO NAVARRETE
GRNC, Parc Científic Universitat de Barcelona & Hospital Sant Joan de Déu
Dept. de Psicologia Bàsica, Universitat de Barcelona

In this article we discuss different views about how information flows through the lexical system in bilingual speech production. In the first part, we focus on some of the experimental evidence often quoted in favor of the parallel activation of the bilinguals' two languages from the semantic system in the course of language production. We argue that such evidence does not require us to embrace the existence of parallel activation of the two languages of a bilingual. In the second part of the article, we discuss the possibility that the language-not-in-use (or the non-response language) is activated via feedback from the sublexical representations and we devise some experimental procedures to assess the validity of such an assumption.

1. Introduction

One of the most remarkable abilities of bilingual speakers is that of being able to choose words from one language while preventing massive interference from the other language. That is, bilingual speakers are able to place themselves in what is called the "monolingual mode" and select representations that belong to only one of their lexicons. In such a scenario, there is one question that models of speech production need to answer: Do the representations of the language-not-in-use affect the production process in the language in use? And if so, how do bilingual speakers prevent massive interference from those representations? In order to answer these questions we need to advance in our knowledge of at least two properties of the functional architecture of the bilingual production system. First, we have to know whether the representations of the language-not-in-use are activated at all during production in the response language. Second, we need to have a clear understanding of the attentional control mechanism that allows speakers to focus on only one set of representations. The first issue is related to the dynamical aspects of speech production and the second to the processes responsible for selecting representations at different levels of processing.

* The preparation of this manuscript was supported by two grants from the Spanish Government (BSO2001 3492-C04-01 BFF2002-1601 10379-E) and by the McDonnell grant "Bridging Mind 1602 Brain and Behavior", and by a grant from the NIH (DC04 542). Albert Costa was supported by the research program "Ramón y Cajal" from the Spanish government. The authors are grateful to Núria Sebastián-Gallés and Mikel Santesteban for their comments. Requests for reprints may be addressed to Albert Costa.

Address for correspondence
Albert Costa, Dept. Psicologia Bàsica, Universitat de Barcelona, P. Vall d'Hebron, 171, 08035 Barcelona, Spain
E-mail: *acosta@ub.edu*

These two aspects are intimately related but we believe they can also be addressed, to some extent, independently. In fact, the type of control mechanisms – specific to cases of bilingualism – that we may need to postulate depends, to a certain degree, on the extent to which the representations of the non-response language are activated in the course of language production. For example, if it turns out that lexical activation is restricted to the representations belonging to the language in which the speaker wants to convey her message, perhaps we do not need to postulate any control mechanism specific to bilingualism that operates over these lexical representations. Thus, before we entertain the type of control mechanisms that operates over the lexical and phonological representations of bilingual speakers, we need to know whether the corresponding representations in the language-not-in-use are activated at all. The present article discusses the theoretical views and empirical data that inform us about the flow of activation through the bilingual lexical systems in the course of language production.

The arguments developed in this article revolve around the following question: What is the contribution to lexical activation of the non-response language from the semantic system and from the phonological system, if any? That is, we will discuss whether in the course of language production: a) the semantic system activates (and to what extent) the lexical representations of the non-response language, and b) the phonological representations of the target language activate (and to what extent) the lexical representations of the non-response language.

The article is organized as follows. First, we present different views of how information flows through the lexical system in monolingual contexts. We will argue that

there is compelling evidence to embrace an interactive model of lexical access. Second, we discuss some of the experimental evidence most often cited in favor of the parallel activation of the bilinguals' two languages from the semantic system (in a feed-forward manner). When doing so we pay special attention to some methodological shortcomings of these experiments and we argue that such evidence is weak and in some cases difficult to interpret because of these methodological shortcomings. Third, we discuss how lexical representations of the non-response language might be activated via feedback from the sublexical representations. Given the lack of experimental evidence in this context, we devise some experimental procedures to assess the validity of this proposal.

2. Functional dynamics in speech production

One of the most influential ideas in cognitive psychology is that of spreading activation among related representations – an idea introduced by Collins and Loftus (1975). According to this principle any representation spreads a proportion of its activation to other representations with which it is linked. In the following, we discuss how (and to what extent) models of speech production have embraced such an assumption.

Models of speech production assume at least three layers of representation: the conceptual, the lexical and the phonological layer. Such models have widely adopted the spreading activation principle when characterizing the flow of activation within the conceptual level. For example, most models agree that when producing the name of a picture ("tiger") there is multiple activation of conceptual information related to such a concept ("dog", "stripes", "lion"). It is also widely assumed that such activation at the conceptual level percolates to the level at which words are represented, the lexical level (e.g. Dell, 1986; Levelt, 1989; Caramazza, 1997).[1] Thus, in the course of naming a picture, there are several lexical representations activated (e.g. "tiger", "dog", "stripes", etc.). Much less agreement is found regarding the functionality of the spreading activation principle between lexical and sublexical (phonological segments) representations.

The so-called discrete models (e.g. Levelt, 1989; Levelt, Schriefers, Vorberg, Meyer, Pechmann and Havinga, 1991; Levelt, Roelofs and Meyer, 1999) restrict the activation flow from the lexical to the sublexical level, in such a way that only one lexical representation (the selected one) activates its phonological content. These models assume that those activated lexical representations that are not selected for production DO NOT SPREAD ACTIVATION to their corresponding phonological properties.

In contrast, the so-called cascade models (Dell, 1986; Starreveld and La Heij, 1995, 1996; Caramazza, 1997; Rapp and Goldrick, 2000; Navarrete and Costa, 2005) assume that any activated lexical representation spreads some activation to its phonological segments. That is, in the course of naming a "tiger" there is sublexical activation of co-activated lexical representations. On this view, spreading activation is a functional principle that characterizes the dynamics of lexical access at all levels of representation.

Thus, these two views of how activation flows from the semantic system to the lexical and sublexical levels make different predictions about what information is activated, and when, at each level of representation.

Up to now we have been concerned with the feed-forward flow of activation in speech production. In fact, all models of lexical access agree that lexicalization starts in a feed-forward manner (from the semantic system): lexical representations are firstly activated from the semantic system. However, one of the most influential proposals also assumes that activation spreads both forwards and backwards (e.g. Dell, 1986; Dell and O'Seaghdha, 1991; Rapp and Goldrick, 2000). According to this view, activation spreads bi-directionally through the different levels of representation implicated in language production. As a consequence of such a bidirectional flow, in the course of naming a "tiger", not only semantically related lexical representations would become activated (e.g. "lion", "dog", "stripes", etc.) but also phonologically related ones (e.g. "tile", "tight"). This is because the activation of the phonological content of the target word (t, ai, g, e, r) sends activation backwards to any lexical representations with which it is connected (e.g. "tile", "tight"). As we will see below, whether or not such a principle is functional in the case of bilingualism has important implications.

The implementation of the spreading activation principle at different levels of representation has consequences for the control/selection mechanisms that need to be postulated at each level of processing. The most immediate one refers to the control exerted at the lexical level (e.g. how the system decides which word needs to be prioritized for further processing). That is, given that there are several lexical representations activated we need to postulate a mechanism that controls (decides) which one should be produced. This mechanism has been labeled LEXICAL SELECTION. In most models, it is assumed that the ease with which the lexical selection mechanism proceeds depends not only on the level of activation of the target lexical representation but also on that of other lexical nodes. Also, if we assume the presence of cascade

[1] Note that there are several proposals about how information flows from the conceptual to the lexical level. For example, Bloem and La Heij (2003) argued that the only conceptual information that sends activation to the lexical level is that included in the preverbal message. However, in doing so, the concept included in the preverbal message activates not only its corresponding lexical node but also those of other semantically related concepts. For our purposes here, all models are equivalent, in the sense that they assume the presence of multiple activation at the lexical level.

processing from the lexical to the sublexical level, we need to postulate a selection mechanism in charge of choosing, among the several segments that are activated, those corresponding to the target word.

Thus, in a model in which activation is allowed to spread freely through the lexical and sublexical levels, we need to postulate selection mechanisms at each of these levels. In contrast, if activation is restricted to the target representation, then the selection process seems quite trivial (e.g. prioritize the only representation that is activated). What about the flow of activation in the bilingual mind?

3. Functional dynamics in bilingual speech production

The critical question regarding the functional dynamics of the bilingual system is the following: Which linguistic representations (e.g. words, phonemes) of the language-not-in-use are activated when bilinguals produce speech in the other language?

There are many occasions in which bilinguals need to restrict their lexicalization to only one language since the use of words from their other language may disrupt communication considerably, given that the interlocutor may not know that language. In such circumstances, and given that the speaker is the one who decides in which language to carry out the communicative act, perhaps the semantic system only activates representations of lexical items in the target language. In such a framework, the bilingual would be functionally equivalent to a monolingual and no control mechanism specific to cases of bilingualism that operates over lexical representations would be needed. This is not to say that she would speak as fluently and accurately as a native speaker, but rather that the same selection mechanisms as those employed by monolinguals would be required for her to produce language. Note that this channeling of activation is possible because, unlike in other domains such as word reading (see Dijkstra and Van Heuven, 2002 for an extensive discussion), the choice of the language in which the message needs to be conveyed depends entirely on the speaker's intention.

However, and despite the obvious benefits of restricting activation to one language, models of bilingual speech production postulate that conceptual representations spread activation to the lexical representations of both languages of a bilingual (Green, 1986, 1998; de Bot, 1992; Poulisse and Bongaerts, 1994; Hermans, Bongaerts, de Bot and Schreuder, 1998; Costa, Miozzo and Caramazza, 1999; Costa, 2005; La Heij, 2005; see Figure 1). Such an assumption has led authors to postulate

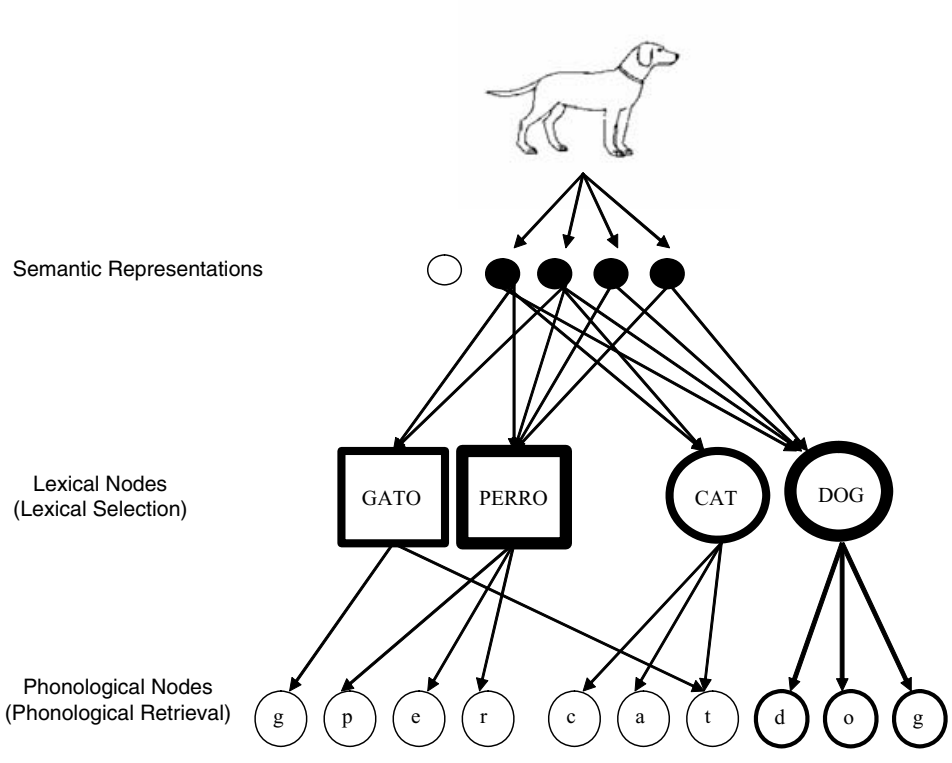

Figure 1. Schematic representation of the activation flow from the semantic to the lexical system of a Spanish–English bilingual speaker in the course of naming the picture of a "dog" in English. The squares represent the lexical nodes of the language not-in-use (Spanish), and the circles represent the lexical nodes of the language in use (English). The arrows represent the flow of activation and the thickness of the circles indicates the level of activation of the representations.

various controlling mechanisms that operate over lexical representations that guarantee that lexical selection is achieved in the intended language, while preventing massive interference from the non-response language. (e.g. Green, 1998; Hermans et al., 1998; Costa and Caramazza, 1999; Lee and Williams, 2001; Costa and Santesteban, 2004a, b).

What about the activation of phonological segments belonging to the lexical items of the non-response language? The majority of the models remain silent about this issue. The only model that specifically claims that the lexical representations of the non-response language also spread some activation to their corresponding phonological segments is that proposed by Costa, Caramazza and Sebastián-Gallés (2000). Bilingual models are also silent about whether activation of phonological segments feeds back to the lexical elements with which they are linked. In other words, they are silent on whether phonological segments activated from the lexical representations of L1 send back some activation to L2 lexical representations (whether interactivity across languages is functional or not).

In the following section, we discuss the empirical basis usually quoted as supporting the notion that the two languages of a bilingual are activated in parallel. We evaluate such results from a critical perspective considering which results actually support such a view and which results are just consistent with it.

4. Feed-forward spreading activation of the two lexicons of a bilingual: Experimental evidence

According to the types of evidence they provide, we classify the findings in support of the parallel activation of the two languages of a bilingual into three main groups: evidence from interference paradigms, evidence from monitoring and evidence from spontaneous slips of the tongue and L1/L2 naming latency.

4.1 Evidence from interference paradigms

Perhaps the most-often cited evidence in support of the co-activation of the two lexicons of a bilingual comes from the presence of cross-language Stroop effects (and Stroop-like effects) in bilingual contexts. Consider, for example, the effects of distractor words on picture naming latencies. In these studies participants have to name a picture while ignoring a distractor word. The best-studied effect is the semantic interference effect: picture naming latencies are higher when the picture (dog) appears along with a semantically (categorically) related distractor ("cat") than with an unrelated distractor ("cap") (e.g. Rosinski, 1977; Lupker, 1979; Glaser and Düngelhoff, 1984). This effect has been interpreted as revealing the larger lexical competition created by a semantically related distractor word in comparison to an unrelated distractor word, and therefore has been used to explore issues related to lexical selection in speech production (e.g. Schriefers, Meyer and Levelt, 1990; Roelofs, 1992; Meyer, 1996; Starreveld and La Heij, 1996; Caramazza and Costa, 2000; Costa and Caramazza, 2002; but see Rosinski, 1977; and Costa, Mahon, Savova and Caramazza, 2003, for an alternative explanation).[2]

In the bilingual version of this paradigm, participants name pictures in one of their languages while distractors are presented in the other language. For example, a Dutch–English bilingual is asked to name the pictures in Dutch (e.g. "hond" ("dog")) while ignoring the presentation of distractor words in English (e.g. "rabbit" or "hammer"). The vast majority of these studies reported semantic interference (e.g. Ehri and Ryan, 1980; Miller and Kroll, 2002; for the Stroop variant of the task see e.g. Preston and Lambert, 1969; Chen and Ho, 1986; Tzelgov, Henik and Leiser, 1990; Altarriba and Mathis, 1997). Following the same rationale used to explain the semantic interference effect in monolingual contexts (see footnote 2), the presence of such an effect across languages is supposed to arise because the activation level of the lexical node of the semantically related distractor word in the non-response language ("rabbit") is larger than that of an unrelated distractor ("hammer"). This differential level of activation comes about because the related lexical node ("rabbit") receives activation from two sources (the presentation of the distractor word "rabbit", and the activation sent by the semantic representation of the target word "hond" ("dog" in Dutch) while the unrelated lexical node ("hammer") receives only activation from one source (the presentation of the distractor word). Thus, the presence of semantic interference across languages is often quoted as supporting the idea that: a) the two lexicons of a bilingual are activated in parallel during the course of language production, and b) the lexical representations of the non-response language compete during the selection of the target representation in the response language. In short, it appears that this effect allows us to "kill two birds with one stone" since it indicates the presence of parallel activation of the two

[2] The logic behind the interpretation of the semantic interference effect in terms of lexical competition is the following. In the course of naming the target picture ("dog"), several semantically related lexical items are activated (e.g. "cat"). The presentation of the distractor word increases the activation of its corresponding lexical node. The activation reached by the lexical node corresponding to a semantically related distractor ("cat") is higher than that of a semantically unrelated distractor ("cap"). This is because the former lexical node ("cat") receives activation from two sources (the presentation of the distractor (cat) and the target's semantic representation ("dog"), while the latter lexical node ("cap") receives activation from only one source (the presentation of the distractor). As a consequence, a semantically related distractor produces more interference than an unrelated distractor.

languages of a bilingual and the presence of cross-language lexical interference.

However, such a conclusion needs to be tempered because of the following three concerns. First, there are alternative explanations of this effect that do not necessarily require these assumptions. In fact, as argued by Costa et al. (1999) the presence of cross-language interference does not necessarily originate because of the competition produced by the lexical representations of the non-response word ("rabbit") in the selection of the target word in the response language ("hond"). Instead, it may be revealing competition within the response language rather than between languages. The argument goes as follows. If the distractor word presented in the non-response language (English: "rabbit") is automatically translated into the response language (Dutch: "konijn") then selection of the target name in the response language ("hond") might be hampered by the high level of activation of the lexical node of the distractor in the response language. That is, it is not the case that "rabbit" competes with the selection of "hond" but rather its translation in the response language, "konijn", does. Under such an explanation, semantic interference across languages does not require that we assume either that there is parallel activation of the two lexicons of a bilingual or that there is lexical competition between the two languages of a bilingual.[3]

Second, several researchers have questioned the notion that the semantic interference effect actually (or at least only) indexes competition between lexical representations (e.g. Rosinski, 1977; Lupker, 1979; Luo, 1999; Costa, Alario and Caramazza, 2005). Instead, they argued that this effect may have its origin at the level at which semantic representations are selected for production. If this interpretation of the phenomenon were to be correct, we could not take these results as indexing lexical competition either within language or across languages. Resolution of this issue requires further research.

A third concern, more important in the present article, is the actual experimental context used in the studies discussed above. The aim of those experiments has been to explore whether the representations of the non-response language are activated in the course of producing the target language. To tackle this question, the use of experimental contexts in which distractor words from the non-response language are presented is far from ideal. This is because the non-response language is already activated from the word recognition system, and hence it is difficult to disentangle such source of activation from that coming from the semantic system. That is, in this experimental context we cannot be sure that the participant is in a purely monolingual mode. So, ideally, one should assess whether there is activation of the non-response language in experimental circumstances in which such a language is not called into play at all (see Grosjean, 1998a, b for a discussion).[4]

Given these problems, we should conclude that while cross-language interference is consistent with the notion of co-activation of the two languages of a bilingual, it does not necessarily require such an assumption.

4.2 Evidence from monitoring

The most compelling evidence supporting the notion of parallel activation of the two languages of a bilingual is that reported by Hermans (2000) and Colomé (2001). In Colomé's study, Catalan-Spanish participants were asked to decide whether a given phoneme was present in the Catalan name of the target picture (see also Wheeldon and Levelt, 1995; and Costa, Sebastián-Gallés, Pallier and Colomé, 2001, for studies in which the phoneme monitoring in speech production has been used). In the critical cases, the target phoneme was present in the Spanish name of the target picture. For example, in some trials participants were asked to decide whether the Catalan name of the target picture "taula" ("table") contained the target phoneme /m/; and in other trials participants were asked to decide whether it contained the target phoneme /f/. Both types of trial required a negative response, since neither /m/ nor /f/ are present in the target word "taula". However, the target phoneme /m/ is present in the Spanish name ("mesa") of the target picture ("table") while the target phoneme /f/ is not. The results showed that it was harder for participants to reject that a given target phoneme was not present in the Catalan

[3] Hermans (2000) has argued that this alternative explanation cannot capture the time-course of the semantic interference effects within and across languages, since both effects have similar time-courses. However, and besides the intrinsic problems when interpreting the time-courses of the effects in this paradigm (see, for example, Starreveld, 2000 and Costa, Colomé, Gómez and Sebastián-Gallés, 2003), a close inspection of the actual results does not seem to be totally consistent with this interpretation. In fact, at negative SOA –300 (e.g. when the word is presented 300 ms before the target picture) within language semantic interference is double in magnitude (44 ms) compared to the between language semantic interference (17 ms). Furthermore, the peak of the semantic interference effect for within- and between-languages condition is placed at different SOAs –300 ms for within- and –150 ms for between-languages). Thus, one may argue that this observation in fact supports a different time-course of the two effects.

[4] Note that this methodological shortcoming applies also to the studies in which the language switching task is used (e.g. Meuter and Allport, 1998, Lee and Williams, 2001; Costa and Santesteban, 2004a). Although these studies may be very useful to understand the control mechanisms used by bilingual speakers during word production, they are not so useful to inform us about whether or not the non-response language is activated in the course of speech production in one language. This is because, arguably, in a language switching task participants may have their two languages active in a way that is not comparable to cases in which they are speaking in only one language.

name of the picture when such a phoneme was present in the Spanish translation than when it was not. This observation was interpreted as revealing that the target's translation was activated in the course of retrieving the picture's name in the target language.

However, this interpretation hinges on the (often implicit) assumption that the presentation of a context stimulus (in this paradigm the target phoneme) has no effect whatsoever on the way the target is processed. This assumption does not need to be correct. It is conceivable that the activated target phoneme feeds back to lexical items in both languages (see our discussion of the role of feedback below), thereby INDUCING the activation of the item in the non-response language rather than REVEALING its activation in normal language production situations.

4.3 Evidence from spontaneous slips of the tongue and L1/L2 naming latencies

One of the few studies that have rigorously assessed the types of slips of the tongue produced by bilingual speakers is that conducted by Poulisse and Bongaerts (1994). In this study, the L1 traces in the slips of the tongue produced in L2 were analyzed. There were two sorts of slips relevant for our purposes here: a) slips in which a word in L1 is produced rather than a word in L2 (lexical intrusions), and b) slips in which an L1 word and its L2 translation are blended. The existence of such errors was interpreted as revealing the fact that the two languages of a bilingual are activated from the semantic system, and that this activation occasionally leads to a misselection of the target word in the response language.

However, although slips of the tongue are an interesting source of information when testing theories of speech production (e.g. Fromkin, 1973; Dell and Reich, 1981), their scope is sometimes rather limited (see Meyer, 1992 for a discussion of this issue). This is especially so in the case of cross-language intrusions.

Language intrusions can certainly be explained in terms of a failure of the lexical selection mechanism, which, instead of picking out the target word from the intended language, selects its translation in the non-response language. Presumably, this failure comes about because of the larger activation level of the target's translation in the non-response language, hence suggesting the parallel activation of the two languages of a bilingual.

However, one cannot take this result as supporting the parallel activation of the two languages of a bilingual. Despite the fact that these observations escape the methodological shortcomings of the previous experiments (since only one language was relevant in the task at hand), they bring another problem. The problem is that language intrusions might have a different origin than a failure in the lexical selection mechanism. In fact, they may stem from an occasional derailment of the general mechanism that channels activation to only one of the lexicons of a bilingual preventing at the same time parallel activation of the two languages. As a consequence of such malfunctioning the two lexicons of a bilingual become activated causing troubles during lexical selection. These troubles could in some cases result in cross-language intrusions and cross-language blends. According to such an explanation, normal "error-free" speech production in bilingual speakers would not entail the co-activation of their two languages. And, in fact, the cross-language errors would be reflecting those cases in which, by mistake, such co-activation is present. Thus, strictly speaking, this sort of evidence is as consistent with the notion that the two languages of a bilingual are activated simultaneously as with the opposite view. At present, it is unclear how one can adjudicate between the two possible origins of these language intrusions (failure in lexical selection vs. failure in channeling activation).

Other evidence that could be understood as supporting the notion of parallel activation of the two languages of a bilingual is the slower naming latencies observed in L2 in comparison to L1 even in proficient bilinguals (e.g. Costa et al., 2000). The argument here is that the lexical representations of L1 would act as more powerful competitors when speaking in L2 than vice versa. As a consequence, L2 naming would be slowed down in comparison to L1 naming. However, there are many other factors that could be behind the L1-L2 naming latencies difference. For example, L1 naming may be faster because, presumably, L1 words have been acquired earlier than their corresponding L2 translations. And given that early-acquired words are named faster than late acquired words (e.g. Morrison, Ellis and Quinlan, 1992; Ellis and Morrison, 1998) one should expect faster naming latencies in L1 than in L2. Along the same lines, L1 words are presumably used more often (and are more familiar) than L2 words and hence their naming advantage. Given these considerations, here again the L1 naming advantage over L2 does not allow us to safely conclude about the existence of co-activation of the two languages of a bilingual.

Another phenomenon that has been interpreted as revealing the presence of activation of the two languages of a bilingual is the cognate advantage observed in several tasks (Roberts and Deslauriers, 1999; Costa et al., 2000; Gollan and Acenas, 2004; Kohnert, 2004). For example, picture naming latencies are lower for pictures whose names are cognates (words whose translations are phonologically similar ["guitar"–"guitarra"]) than for pictures whose names are non-cognates (words whose translations are phonologically dissimilar ["car"–"coche"]). According to the interpretation given to this effect by Costa et al. (2000), the advantage of cognates over non-cognates arises because of a facilitation in the retrieval of the phonological content of the target

word. The argument goes as follows. If in the course of naming an object the two languages of a bilingual are activated, and activation from both languages percolates to the phonological level, then the phonological representation of a cognate word will receive activation from two sources (the target lexical node ["guitar"] and its translation ["guitarra"]) while the phonological representation of a non-cognate word will receive activation from only one source (the target lexical node). As a consequence, the retrieval of the phonological properties of cognate words would be easier/faster than those of non-cognate words. According to such an explanation, the cognate effect would be revealing: a) the co-activation of the lexical representations of the two lexicons from the semantic system, and b) the activation of the sublexical properties of the lexical representations of the non-response language.

However, this cognate effect may have a different origin. For example, some researchers argued that cognate translations are semantically more similar than non-cognate translations (Van Hell and de Groot, 1998). If that were to be the case, one may claim that cognate words share their conceptual representation while non-cognate words do not. In this context, and under the assumption that the ease with which a semantic representation is retrieved depends, among other things, on the frequency of retrieval/usage, it is possible that the semantic representations corresponding to cognate words are retrieved, everything else being equal, faster than those corresponding to non-cognate words, and hence the advantage of cognates in naming latencies.

There is also a second interpretation of the cognate effect that does not require the assumption of parallel activation of the two lexicons of a bilingual. Kirsner, Lalor and Hird (1993) argued that cognate translations might share their morphological stem (even in the case of monomorphemic words), while non-cognates do not. In such view, the retrieval of the morphological stem of cognate words is faster because shared morphemes would be more frequent, everything else being equal, than non-shared morphemes.

A third possible origin of the advantage of cognates over non-cognates relates to differences between the learning difficulties of these two types of words. Arguably cognates, because of their obvious similarity with the L1 translations, are learned more easily and more robustly than non-cognates (and even they are used more often when two alternatives are available). For example, a Spanish-English unbalanced bilingual may be more inclined to use the word "construct" ("construir" in Spanish) than the word "build" ("construer" in Spanish). If this learning advantage were to have future consequences for language processing (e.g. see the persistent effects of age of acquisition in adulthood Ellis and Morrison, 1998; Hodgson and Ellis, 1998) then cognates may enjoy a processing advantage over non-cognates in adult language production.

Thus, although we believe that the interpretation of cognate effects in terms of facilitation in the retrieval of phonological information is the appropriate one (see for a discussion Costa, Santesteban and Caño, 2005), at present such an effect is also consistent with models that do not assume the presence of co-activation of the two languages of a bilingual. More research is needed in order to determine the precise origin(s) of the cognate effect.

5. Interim summary

We have argued that the experimental evidence most-often quoted as supporting the parallel activation of the two languages of a bilingual is not conclusive, since there are either alternative explanations that do not require such a parallel activation, or there are important methodological shortcomings that prevent us from definitely embracing such an assumption. The most important methodological shortcoming concerns situations in which (a) the two languages of a bilingual are called into play during the production task, as is the case with the picture-word interference paradigm and the switching paradigm (see footnote 4) and (b) the presentation of context stimuli may invoke the activation of items in the non-response language, as may happen in the phoneme monitoring task. If we want to know whether in a common monolingual production setting the two languages of a bilingual become activated in parallel, we need to restrict our experimental situation to only one language. Because of these reasons, we believe that to draw a definite conclusion is not totally justified. Given the crucial implications of embracing such an assumption for models of bilingual speech production, it is reasonable that we step back and design new experiments that seek to find a resolution of this issue. This is an important enterprise because if it turns out that bilinguals can channel the activation coming from the semantic system according to the target language, then the subsequent cognitive processes involved in speech production would be functionally similar to those of monolingual speakers, at least for highly-proficient bilinguals. In assessing this issue it is critical that we use paradigms in which only the representations of one of the lexicons of a bilingual are relevant for the task at hand (e.g. picture naming in only one language).

Up to here we have discussed the experimental evidence cited in favor of concurrent activation of the two languages of a bilingual from the semantic system (feed-forward spreading activation from the conceptual system to the two languages of a bilingual). That is, we have evaluated the evidence that speaks to the first part of the question presented in the Introduction: What is the contribution to lexical activation in the non-response language from a) the semantic system and b)

the phonological system, if any? In the following we will address the second part of the question.

As we argued, the lexical representations of the non-response language may become activated via the activation of the phonological segments of the target word in the response language. That is, if the interactivity assumption is correct (and functional across languages) lexical representations from the non-response language may become activated from the phonological level. Given the lack of empirical evidence on this issue, in the next section we discuss various ways in which we can test the extent to which lexical representations of the non-response language may become activated via-feedback from the phonological system.

6. Backward spreading activation to the two lexicons of a bilingual: Searching for evidence

As discussed in the Introduction, there are some models of lexical access that assume a bi-directional flow of activation in the course of lexicalization. As in any other model, lexicalization starts in a feed-forward manner from the semantic system to the lexical and sublexical systems. However, before selection of the target word is achieved, the activation of the sublexical representations bounces back to the lexical representations with which they are linked. Although a review of the evidence supporting interactive models falls outside of the scope of the present article, it is nevertheless important to stress that in the last years reaction time, slips of the tongue and aphasic studies have provided compelling evidence supporting such a principle (e.g. Cutting and Ferreira, 1999; Foygel and Dell, 2000; Goldrick and Rapp, 2002; Gordon, 2002; Ferreira and Griffin, 2003; Stemberger, 2004). In fact, interactivity (in some way or another) is currently assumed by the majority of the models. Thus, in such a context, it is pertinent to ask whether it is also functional across the two languages of a bilingual. That is, we should address the question of whether the activation of the sublexical properties of the target word in the response language sends activation back to the lexical representations of only the response language or also to those of the non-response language. Thus, the issue at stake here is the language-specific or non-specific nature of the interactive principle in cases of bilingualism.

It is important to note that the issue of interactivity across languages is independent of whether the semantic system activates the two languages in parallel. In fact, there are several ways in which these two assumptions may combine. It is possible that the semantic system activates only the lexical representations of the response language and nevertheless its phonological representations activate the lexical representations (via feedback) of the non-response language. At the same time, the semantic system may activate both lexicons in parallel and interactivity across languages may be absent. Thus, in principle, we could evaluate the different proposals independently, and importantly none of the effects reported above are problematic for an interactive model.

6.1 Neighborhood effects across languages

Neighborhood density refers to how many phonologically similar words a given word has. The way in which density is calculated is by changing one phoneme at a time of a word in a given language and counting how many of the resulting items are existing words in that language. For example, a word such as "cat" in English has a dense neighborhood because many of the resulting items when changing a single phoneme are real words (e.g. pat, that, mat, chat, sat, cut, cot) whereas a word such as "cry" has a sparse neighborhood (e.g. fry, try, dry).

Interestingly, neighborhood density correlates negatively with naming latencies (responses are faster for pictures that have names with dense neighborhoods than for pictures that have names with sparse neighborhoods, everything else being equal; Vitevitch, 2002). Also, words with dense neighborhoods are less vulnerable to falling in tip of the tongue states than words with sparse neighborhoods (Harley and Brown, 1998; Vitevitch and Sommers, 2003). Neighborhood density also predicts, to some extent, successful speech production of both normal (Vitevitch, 1997; Stemberger, 2004) and aphasic speakers (Gordon, 2002).

The most widespread interpretation of the neighborhood effect assumes that the effect arises as a consequence of the interactive nature of the speech production system. The argument goes as follows. In the course of naming a picture ("cat"), the phonemes of such a word are activated (/k/, /æ/, /t/). Before selection takes place, those phonemes send activation back to all words with which they are connected ("rat", "mat", "cot" etc.), which in turn send activation forward to their corresponding phonemes. If a word has many neighbors, there will be many words sending activation to the shared phonological content, which in turn will send much activation to the target word. Hence, the retrieval of both the lexical node of the target word and of its phonemes would be facilitated because of the multiple sources of activation. In contrast, when a word has few neighbors ("cry") such a benefit will be smaller.

How can this effect inform us about the co-activation of the two languages of a bilingual? If we assume that: a) phonemes or phonological features are partially shared across languages, and b) interactivity is functional across languages; then we should expect the lexical nodes of the non-response language to be active in the course of language production. This is because the phonological properties of the target word in the response language

would send back activation to any lexical node with which they are linked, activating therefore words in the response and in the non-response language. For example, if a Spanish-English bilingual is naming the picture of a "can" in English, the activation of the phonemes /k/, /æ/ and /n/ would send activation not only to the English words "man", "ran" and so on, but also to Spanish words such as "pan", "con", "cal".[5]

In such a scenario, one may predict neighborhood density effects across languages. That is, one could assess whether the best predictor of naming latencies is the neighborhood density value resulting from considering only the words in the response language (the language specific neighborhood), or that resulting when considering both the words in the response and in the non-response language (the language non-specific neighborhood). Ideally, one should look at words that have a sparse neighborhood in the response language and a dense neighborhood in the other language. To illustrate, consider the case of a monosyllabic word in Spanish such as "mil" (thousand). This word has few neighbors in Spanish (vil, mal, mis), however it has many neighbors in English (e.g. kill, chill, gill, bill, till, miss). Thus, the question here is whether the word "mil" when produced in Spanish behaves as a word of a dense or of a sparse neighborhood. If the non-response language (English) is not activated in the course of naming in the response language (Spanish), then naming latencies for "mil" should resemble those of another control word with a sparse neighborhood in both languages, everything else being equal. However, if the two languages are activated then such a word should behave as any other word with a dense neighborhood. Thus, exploring the presence of neighborhood effects across languages could be a window to assess whether there is activation of the two lexicons of a bilingual. Of course, when doing so one needs to consider the relative dominance of the bilingual speaker, since the size of the L2 neighborhood will certainly depend on the L2 knowledge of the speaker. Perhaps a way to partially circumvent this problem would be to test participants in their L2, and evaluate the effects of the L1 neighborhood that is presumably comparable across speakers.

6.2 Phonologically similar words without semantic overlap: False friends

Another way to explore the existence of co-activation of the non-response language is to assess the effects of words that are very similar in phonological form across languages but completely different in meaning. These items are often referred to with the terms "false friends" or "quasi-homophones across languages". For example, words such as "net" ["clean" in Catalan], "sock" ["I am" in Catalan], "pot" ["S/he can" in Catalan], "lot" ["flashlight" in Catalan], "cop" ["blow" in Catalan], or "pet" ["fart" in Catalan] are phonologically very similar in English and Catalan but mean very different things in the two languages. If interactivity is functional across languages, we should expect that in the course of naming a word in one language its corresponding "false friend" in the non-response language is activated. That is, when naming the picture of a net in English, the lexical node "net" ("clean") in Catalan will become activated. This is because the phonological properties of the English word (/n/, /e/, /t/) will send activation back to any lexical item with which they are linked, and, given the large overlap of the phonological segments between false friends, the word in the non-response language will be activated as well. As a consequence of this concurrent activation, and everything else being equal, the phonological representation of the words in the response language will be highly available in the course of naming in comparison to that of other words that do not have false friends (see Figure 2). This is because, in the case of false friends, the phonological properties of the target word in the response language (e.g. net in English) will receive activation from two sources, the target word in the response language and the false friend in the non-response language (e.g. "net" in Catalan ["clean"]). Note that if this prediction were to be borne out by the experimental results, it would be hard for models that do not assume interactivity (regardless of whether they assume parallel activation of the two languages of a bilingual) to account for them (see Kroll, Dijkstra, Janssen, and Schriefers, 2000; for preliminary results on this issue). The only way in which a false friend can be activated in the course of producing a word in the other language is through the activation sent by the phonological representation of the target word to the lexical representation of the false friend in the non-response language.

Furthermore, false friends also offer us the possibility of testing whether the semantic system activates the two languages of a bilingual. Before going into this issue, it is worth introducing the study conducted by Cutting and Ferreira (1999), in which participants were asked to name pictures with homophone names while ignoring distractor words. For example, they were asked to name the picture of a "ball (toy)", and were presented with distractors that were semantically related to the meaning of the target's homophone twin (e.g. "dance") or with unrelated distractors (e.g. "hammer"). Distractors semantically related to the non-depicted meaning of the picture name ("dance") led to faster responses than unrelated distractors ("hammer"). According to the

[5] For this argument to work, it is not necessary that we assume that the phonemes of the two languages are shared, but just that we assume that at some level of representation (e.g. phonological features such as plosive) there is certain overlap between languages. The fact that bilingual speakers very often show a foreign accent is consistent with this idea.

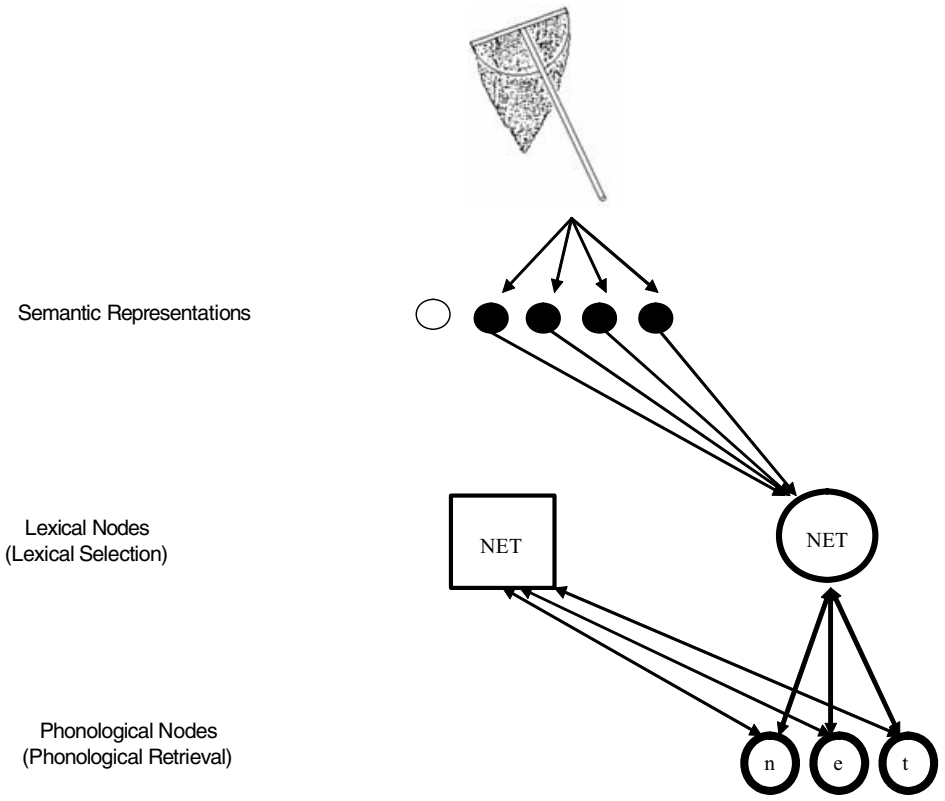

Figure 2. Schematic representation of the lexical system of a Catalan–English bilingual speaker in the course of naming a picture in English whose name "net" has a false friend in Catalan ("net" in Catalan means "clean"). The squares represent the lexical nodes of the language not-in-use (Catalan), and the circles represent the lexical nodes of the language in use (English). The arrows represent the flow of activation and the thickness of the circles indicates the level of activation of the representations. Note that the lexical representation of the Catalan false friend is activated from the phonological properties of the target word in the response language (English).

authors, this effect comes about because: a) the distractor word "dance" activates its semantic representation along with a cohort of semantically related concepts that include "ball (dance)", b) the activation of the concept "ball (dance)" spreads in a cascade fashion to the lexical and sublexical systems, activating the same phonological content of the target picture's name "ball (toy)". In such circumstances, the retrieval of the phonological content of the target name is sped up. Given that false friends could be considered as quasi-homophones across languages, one could ask the question of whether a distractor word semantically related to the meaning of the false friend in the non-response language would affect naming latencies. For example, a Catalan-English bilingual is asked to name the picture of a "net" in English while ignoring the presentation of a distractor word in English that happens to be the translation of the Catalan meaning of the word "net (clean)". If in the course of language production the two languages of a bilingual are activated from the semantic system, one should expect that the conceptual representation of "clean" activates not only its lexical representation in the response language ("clean") but also in the non-response language ("net"). And, following the same reasoning of Cutting and Ferreira (1999), we should observe a priming effect. That is, responses would be faster when the distractor word corresponds to the translation of the meaning of the false friend (e.g. "clean") than when it is unrelated (e.g. "quick") (see Costa et al., 1999 for a similar rationale).[6]

Thus, a closer look at how false friends are processed could help us to asses not only the presence of interactivity across languages but also of co-activation of the two languages of a bilingual from the semantic system.

[6] Costa et al. (1999) conducted an experiment similar to the one proposed here in which a distractor word in the non-response language – Spanish – ("pelea (fight)") had a translation in the response language – Catalan – ("baralla") that was phonologically similar to the target word in the response language – Catalan – ("baldufa (spinning top)"). In principle, if the distractor word activates its translation one could expect to find a phonological effect, given the overlap between that translation and the target. However, we did not find any phonological effect. Note, however, that the phonological overlap in case of the homophones used by Cutting and Ferreira (1999) and in the case of false friends is much larger.

6.3 Language constraints in slips of the tongue

Although at first sight slips of the tongue do not seem to follow any systematic pattern, they actually do. One of the best-studied phenomena in the error literature is the lexical bias effect (LBE). This effect refers to the fact that phonological slips of the tongue (errors involving the misselection of a phonological element) tend to result in existing words more often than what one would expect by chance (e.g. Baars, Motley and MacKay, 1975; Dell and Reich, 1981; Dell, 1986). That is, if the speaker makes a phonological error in the production of, say, the word "cat", it is more likely that she will end up producing an existing word such as "mat" than a non-word such as "yat".

The LBE finds a ready explanation in an interactive framework. If, due to a momentarily malfunctioning of the phonological encoding process, a phoneme is replaced by another, the most probable intruder would be the phoneme with the highest level of activation. Importantly, in an interactive system, the level of activation of a given phoneme would depend, among other things, on whether or not it receives activation from an existing word in the lexicon. Thus, the intruder would tend to be a phoneme of an existing word and hence the result of the intrusion would also be an existing word. Consider the example of producing "mat" instead of "cat". At the moment of selecting the phoneme /k/ the phoneme /m/ would be more activated than the phoneme /y/, because /m/ would be receiving activation from the word "mat", while the phoneme /y/ will not be receiving activation at all. Thus, if there is a misselection when retrieving /k/, the most likely intruder would be /m/ and not /y/, leading to the production of the unintended[7] word "mat".

One way to test whether there is concurrent activation of two languages is to explore whether there is a LBE across languages. Let us discuss how one can approach this issue with an example. The LBE has been studied by using error-elicitation techniques in which sound exchanges are primed. For example, in these experiments the participant is asked to repeat pairs of words in which each element of the pair starts with a specific consonant (e.g. "**c**at–**p**in", "**c**ot–**p**ew", "**c**ar–**p**en", "**c**ouch–**p**ig"). At the end of this list the participant is asked to repeat a target pair in which the same critical consonants are swapped ("**p**an–**c**ot"). Crucially, there are two sorts of target pairs: a) pairs in which an exchange error involving the two first consonants leads to two existing words ("pan–cot" would result in "can–pot"), and b) pairs in which an exchange error involving the two first consonants leads to two non-words ("pal–cute" would result in "cal–pute"). The standard result is that participants produce more errors in the former case than in the latter (Baars et al., 1975; Hartsuiker, Corley and Martensen, 2005). That is, participants tend to swap the first phoneme of each word more often when the result of such an exchange leads to two existing words than when it does not. We can adapt this experimental setting to assess whether the non-response language affects the presence of a LBE. We could design an experiment with word pairs such that, when the first sound is switched, this results in two non-words in the response language, but which are at the same time two existing words in the non-response language. That is, an error in the target pair would result for some pairs in two words in the non-response language and for other pairs in two non-words. To illustrate, if the first sound of the two words of the Catalan pair "cop–tot" ("blow–all") swap, the outcome would be two non-words in Catalan ("top–cot"), and at the same time it would lead to two existing words in English. In contrast, if the first sound of the two words of the Catalan pair "nit–ruc" ("night–donkey") swap, the outcome would be two non-words, both in Catalan and in English ("rit–nuc"). If Catalan-English bilinguals make more exchange errors in the former type of pairs than in the latter, then we should conclude that the presence of LBE is affected by the lexical items of the non-response language. Such a result would indicate the presence of interactivity between the two languages of a bilingual.

7. Activation from semantics and/or from lexical forms

In the previous sections we discussed: a) the experimental evidence that speaks the concurrent activation of the two languages of a bilingual and b) some experimental proposals to further explore such a question. However, there is a fundamental difference between the studies reviewed in the fourth section and the ones proposed in the sixth section, concerning the source of activation of the non-response language: whether it is activated from the semantic system or from the phonological system (with the exception of the study of false friends that could inform us about the two issues). It is important to recall that these two sources of activation are, to some extent, independent, in the sense that the presence of one does not preclude (or force) the existence of the other. As a consequence, if indeed there is activation of the non-response language, it could stem from the semantic system, the phonological system or from both. However, the sorts of representations that would be activated will be different depending on which combination of these

[7] Another account of the lexical bias effect in which the effect emerges without the need of interactive processing has been put forward among others by Levelt et al. (1999). According to such a view, the lexical bias effect would be the result of a bias in the monitor system that inspects the verbal output before articulation. However, recent results by Hartsuiker, Corley and Martensen, 2005, revealed that although the monitor system may have some role in the presence of the lexical bias, interactivity between the phonological and lexical level is still needed in order to account for such an effect.

assumptions turns out to be the correct one. If co-activation comes ONLY from the semantic system, then semantically related words both in the response and in the non-response language will become activated. However, if activation comes from the phonological level, then we should expect to find activation of lexical representations that are phonologically similar to the target word in the response language. As a consequence, the control mechanism – specific to bilingualism – that we may need to postulate to account for bilingual language production will depend on which of these combinations of assumptions turns out to be the correct one.

8. Potential variables that may affect the dynamics of lexical access in bilingual production

We have assessed the issue of the co-activation of the two languages of the bilingual, in the context in which the bilingual speaker is placed in a monolingual mode. Such situations are not uncommon, since there may be many cases in which the interlocutor does not know the other language of the bilingual speaker.

However, there are several other contexts in which one could posit the same questions we asked here, and for which the same answer is not necessarily granted (see, for example, a discussion of this issue in Grosjean, 1998a, b). For example, one may posit the question of parallel activation of the two languages of a bilingual in contexts in which codeswitching is allowed, or in which the conversation is explicitly conducted in the two languages of an individual. It is entirely possible that these different contexts may lead to different answers to the questions addressed here. That is, one may argue that the cognitive system of the bilingual speaker adapts to the task demands (or production contexts) in which is placed in order to achieve an optimal performance (see Green, 2002 and Thomas, 2002 for a discussion of the adaptative nature of the bilingual system). However, the extent to which the bilingual cognitive system is flexible enough to modulate the activation of the two lexicons of a bilingual during speech production in accordance to the context is an open question that requires further research. Nevertheless, what is important for our purposes here is that if we were to conclude that there actually is activation of the non-response language in a monolingual context, we could conclude that in the other cases (e.g. "bilingual mode contexts") such activation would very likely be present as well.

There are also other properties of the bilingual system that we have not addressed in the present article and that may modulate the extent to which there is activation of the two languages of a bilingual even in monolingual contexts. For example, variables such as the similarity of the two languages of a bilingual, the age (and manner) at which L2 has been acquired, the proficiency achieved in L2, the recency and frequency of use of the two languages, and the discourse topic may affect whether or not the two languages become activated in parallel even in monolingual contexts. An example of how these variables may play an important role on the type of processes involved in speech production can be found in the theoretical proposal put forward recently by Costa and Santesteban (2004a). These authors argue that the control mechanism that guarantees lexical selection in the target language crucially depends on the L2 proficiency of the bilingual speakers. They further argue that there is a qualitative shift from a reliance on inhibitory control to a reliance on language-specific selection mechanisms that is intimately tied to an increase in the L2 proficiency.

Thus, future research is needed to address the contribution of all these variables, along with how they may interact with the production contexts mentioned above, to the presence of activation of the non-response language during speech production.

9. Final remarks

In this article, we have focused on the extent to which there is concurrent activation of the two languages of a bilingual in the course of language production in monolingual contexts. We argued that, despite the widespread assumption that the semantic system activates the two lexicons of a bilingual, the available experimental evidence in support of such a claim is non-conclusive. This is because most of the phenomena reported in the bilingual speech production literature, although consistent with the notion of concurrent activation of the two languages of a bilingual, do find ready alternative explanations that do not require entertaining such an assumption. Furthermore, some of those studies present methodological shortcomings that prevent a definite conclusion. As a consequence, we believe it is worthwhile to take a step back and design new experiments that allow for a clearer answer. In the second half of the article, and given the increasing evidence supporting the interactive nature of the speech production system, we described some ways of assessing the extent to which concurrent activation of the two languages of a bilingual may originate from a language-non-selective interactivity. Resolution of the issue of whether or not the representations of the language-not-in use are activated and the source of this activation is crucial for advancing our knowledge of the dynamics of lexical access in bilingual speech production.

References

Altarriba, J. & Mathis, K. (1997). Conceptual and lexical development in second language acquisition. *Journal of Memory and Language*, 36 (4), 550–568.

Baars, B. J., Motley, M. T. & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of

tongue. *Journal of Verbal Learning and Verbal Behavior*, 14 (4), 382–391.

Bloem, I. & La Heij, W. (2003). Semantic facilitation and semantic interference in word translation: Implications for models of lexical access in language production. *Journal of Memory and Language*, 48 (3), 468–488.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, 14 (1), 177–208.

Caramazza, A. & Costa, A. (2000). The semantic interference effect in the picture–word interference paradigm: Does the response set matter? *Cognition*, 75, B51–B64

Chen, H. C. & Ho, C. (1986). Development of Stroop interference in Chinese–English bilinguals. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 12 (3), 397–401.

Collins, A. M. & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82 (6), 407–428.

Colomé, A. (2001). Lexical activation in bilinguals' speech production: Language-specific or language-independent? *Journal of Memory and Language*, 45 (4), 721–736.

Costa, A. (2005). Lexical access in bilingual production. In Kroll & De Groot (eds.), pp. 308–325.

Costa, A., Alario, F.-X. & Caramazza, A. (2005). On the categorical nature of the semantic interference effect in the picture–word interference paradigm. *Psychonomic Bulletin and Review*, 12 (1), 125–131.

Costa, A. & Caramazza, A. (1999). Is lexical selection in bilingual speech production language-specific? Further evidence from Spanish–English and English–Spanish bilinguals. *Bilingualism: Language and Cognition*, 2 (3), 231–244.

Costa, A. & Caramazza, A. (2002). The production of noun phrases in English and Spanish: Implications for the scope of phonological encoding in speech production. *Journal of Memory and Language*, 46 (1), 178–198.

Costa, A., Caramazza, A. & Sebastián-Gallés, N. (2000). The cognate facilitation effect: Implications for the models of lexical access. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26 (5), 1283–1296.

Costa, A., Colomé, A., Gómez, O. & Sebastián-Gallés, N. (2003) Another look at cross-language competition in bilingual speech production: Lexical and phonological factors. *Bilingualism: Language and Cognition*, 6 (3), 167–179.

Costa, A., Mahon, B., Savova, V. & Caramazza, A. (2003). Level of categorization effect: A novel effect in the picture–word interference paradigm. *Language and Cognitive Processes*, 18 (2), 205–233.

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 (3), 365–397.

Costa, A. & Santesteban, M. (2004a). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, 50 (4), 491–511.

Costa, A. & Santesteban, M. (2004b). Bilingual word perception and production: Two sides of the same coin? *Trends in Cognitive Sciences*, 8 (6), 253–253.

Costa, A., Santesteban, M. & Caño, A. (2005). On the facilitatory effects of cognate words in bilingual speech production. *Brain and Language*, 94, 94–103.

Costa, A., Sebastián-Gallés, N., Pallier, C. & Colomé, A. (2001). El desarrollo temporal de la codificación fonológica: ¿Un procesamiento estrictamente serial? [The time course of segment-to-frame association in phonological encoding: a strictly serial processing?]. *Cognitiva*, 13 (1), 3–34.

Cutting, J. C. & Ferreira, V. S. (1999). Semantic and phonological information flow in the production lexicon. *Journal of Experimental Psychology: Learning Memory and Cognition*, 25 (2), 318–344.

de Bot, K. (1992). A bilingual production model: Levelt's speaking model adapted. *Applied Linguistics*, 13, 1–24.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93 (3), 283–321.

Dell, G. S. & O'Seaghdha, P. G. (1991). Mediated and convergent lexical priming in language production – A comment on Levelt et al. (1991). *Psychological Review*, 98 (4), 604–614.

Dell, G. S. & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, 20 (6), 611–629.

Dijkstra, T. & Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5 (3), 175–197.

Ellis, A. W. & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24 (2), 515–523.

Ehri, L. C. & Ryan, E. B. (1980). Performance of bilinguals in a picture–word interference task. *Journal of Psycholinguistic Research*, 9 (3), 285–302.

Ferreira, V. S. & Griffin, Z. M. (2003). Phonological influences on lexical (mis)selection. *Psychological Science*, 14 (1), 86–90.

Foygel, D. & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, 43 (2), 182–216.

Fromkin, V. (1973). *Speech errors as linguistic evidence*. The Hague: Mouton.

Glaser, W. R. & Düngelhoff, F.-J. (1984). The time course of picture–word interference. *Journal of Experimental Psychology: Human Perception and Performance*, 10, 640–654.

Gollan, T. H. & Acenas, L. A. (2004). What is a TOT? Cognate and translation effects on tip-of-the-tongue states in Spanish–English and Tagalog–English bilinguals. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30 (1), 246–269.

Goldrick, M. & Rapp, B. (2002). A restricted interaction account (RIA) of spoken word production: The best of both worlds. *Aphasiology*, 16 (1–2), 20–55.

Gordon, J. K. (2002). Phonological neighborhood effects in aphasic speech errors: Spontaneous and structured contexts. *Brain and Language*, 82 (2), 113–145.

Green, D. W. (1986). Control, activation and resource: A framework and a model for the control of speech in bilinguals. *Brain and Language*, 27 (2), 210–223.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, 1 (2), 67–81.

Green, D. W. (2002). The bilingual as an adaptive system. *Bilingualism: Language and Cognition*, 5, 206–208.

Grosjean, F. (1998a). Transfer and language mode. *Bilingualism: Language and Cognition*, 1 (3), 175–176.

Grosjean, F. (1998b). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition*, 1 (2), 131–149.

Harley, T. A. & Brown, H. E. (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, 89 (1), 151–174.

Hartsuiker, R. J., Corley, M. & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related reply to Baars, Motley, and MacKay (1975). *Journal of Memory and Language*, 52, 58–70

Hermans, D. (2000). Word production in a foreign language. Ph.D. thesis, University of Nijmegen.

Hermans, D., Bongaerts, T., de Bot, K. & Schreuder, R. (1998). Producing words in a foreign language: can speakers prevent interference from their first language? *Bilingualism: Language and Cognition*, 1 (3), 213–230.

Hodgson, C. & Ellis, A. W. (1998). Last in, first to go: Age of acquisition and naming in the elderly. *Brain and Language*, 64, 146–163.

Kirsner, K., Lalor, E. & Hird, K. (1993). The bilingual lexicon: Exercise, meaning and morphology. In R. Schreuder & B. Weltens (eds.), *The bilingual lexicon: Studies in bilingualism series*, pp. 215–248. Amsterdam: John Benjamins.

Kohnert, K. (2004). Cognitive and cognate-based treatments for bilingual aphasia: A case study. *Brain and Language*, 91 (3), 294–302.

Kroll, J. F. & De Groot, A. M. B. (eds.) (2005). *Handbook of bilingualism: Psycholinguistic approaches*. Oxford: Oxford University Press.

Kroll, J. F., Dijkstra, A., Janssen, N. & Schriefers, H. (2000). Selecting the language in which to speak: Experiments on lexical access in bilingual production. Paper presented at the 41st Annual Meeting of the Psychonomic Society, New Orleans, LA.

La Heij, W. (2005). Monolingual and bilingual lexical access in speech production: Issues and models. In Kroll & De Groot (eds.), pp. 289–307.

Lee, M. W. & Williams, J. N. (2001). Lexical access in spoken word production by bilinguals: Evidence from the semantic competitor priming paradigm. *Bilingualism: Language and Cognition*, 4 (3), 233–248.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J. M., Roelofs, A. & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral & Brain Sciences*, 22 (1), 1–75.

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T. & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, 98 (1), 122–142.

Luo, C. R. (1999). Semantic competition as the basis of Stroop interference: Evidence from color-word matching tasks. *Psychological Science*, 10 (1), 35–40.

Lupker, S. J. (1979). The semantic nature of response competition in the picture–word interference task. *Memory & Cognition*, 7 (6), 485–495.

Meuter, R. F. I. & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language*, 40, 25–40.

Meyer, A. S. (1992). Investigation of phonological encoding through speech error analyses: achievements, limitations, and alternatives. *Cognition*, 42 (1–3), 181–211.

Meyer, A. S. (1996). Lexical access in phrase and sentence production: Results from picture–word interference experiments. *Journal of Memory and Language*, 35 (4), 477–496.

Miller, N. A. & Kroll, J. F. (2002). Stroop effects in bilingual translation. *Memory & Cognition*, 30 (4), 614–628.

Morrison, C. M., Ellis, A. W. & Quinlan, P. T. (1992). Age of acquisition, not word-frequency, affects object naming, not object recognition. *Memory and Cognition*, 20 (6), 705–714.

Navarrete, E. & Costa, A. (2005) Phonological activation of ignored pictures: Further evidence for a cascade model of lexical access. *Journal of Memory and Language*, 53, 359–377.

Poulisse, N. & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics*, 15, 36–57.

Preston, M. S. & Lambert, W. E. (1969). Interlingual interference in a bilingual version of the Stroop color-word task. *Journal of Verbal Learning and Verbal Behavior*, 8 (2), 295–301.

Rapp, B. & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, 107 (3), 460–499.

Roberts, P. & Deslauriers, L. (1999). Picture naming of cognate and non-cognate nouns in bilingual aphasia. *Journal of Communication Disorders*, 32 (1), 1–23.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, 42 (1–3), 107–142.

Rosinski, R. R. (1977). Picture–word interference is semantically based. *Child Development*, 48 (2), 643–647.

Schriefers, H., Meyer, A. S. & Levelt, W. J. M. (1990). Exploring the time course of lexical access in production: Picture–word interference studies. *Journal of Memory and Language*, 29 (1), 86–102.

Starreveld, P. A. (2000). On the interpretation of onsets of auditory context effects in word production. *Journal of Memory and Language*, 42, 497–525.

Starreveld, P. A. & La Heij, W. (1995) Semantic interference, orthographic facilitation and their interaction in naming tasks. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21 (3), 686–98.

Starreveld, P. A. & La Heij, W. (1996). Time-course analysis of semantic and orthographic context effects in picture naming. *Journal of Experimental Psychology: Learning,. Memory and Cognition*, 22 (4), 96–918.

Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, 90 (1–3), 413–422.

Thomas, M. S. C. (2002). Theories that develop. *Bilingualism: Language and Cognition*, 5, 216–217.

Tzelgov, J., Henik, A. & Leiser, D. (1990). Controlling Stroop interference: Evidence from a bilingual task. *Journal of Experimental Psychology: Learning, Memory,. and Cognition*, 16 (5), 760–771.

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

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech*, 40 (3), 211–228.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning,. Memory and Cognition*, 28 (4), 735–747.

Vitevitch, M. S. & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory and Cognition*, 31 (4), 491–504.

Wheeldon, L. R. & Levelt, W. J. M. (1995). Monitoring the time-course of phonological encoding. *Journal of Memory and Language*, 34 (3), 311–334.

Received December 9, 2004
Revision received June 18, 2005
Accepted August 1, 2005