This article reviews representative bilingual speech production models after describing three contrastive notions pointed out by Costa (2005) and La Heiji (2005). The three key notions that are to be described first are: 1) language specific versus language nonspecific activation; 2) language specific versus language non-specific selection; 3) simple access, complex selection versus complex access, simple selection models. Subsequently, major bilingual speech production models are reviewed: De Bot (1992); De Bot and Schreuder (1993); Poulisse and Bongaerts (1994); Green (1998a, b); Kormos (2006). This article concludes with discussions of three issues in bilingual speech production studies that are to be pursued in future research: locus of language selection, inhibition, and developmental perspectives of speech production models.

Critical notions in the study of bilingual lexical selection

Language specific versus language nonspecific activation

The issue of language specific versus language nonspecific activation of
lexical access concerns whether words in the non-intended language are co-activated during L2 speaking. To put this in a different way, as Costa (2005) states, language specific activation flow means that activation of lexical items is restricted only to one language, which creates similar situations to activation patterns of L1 speech production. Straight to the point, a general consensus is reached over this issue; thus, words in the non-intended language are also active during L2 speaking (e.g., Costa, 2005; Kormos, 2006; La Heij, 2005; Poulisse & Bongaerts, 1994; cf. Lee & Williams, 2001). An oft-cited study to support this claim is Hermans, Bongaerts, De Bot, and Schreuder (1998).

Hermans et al. (1998) used a picture-word interference task to investigate the activity of the non-target language during the target language production. One of the research questions was whether the L1 word (Dutch) was activated during L2 word production (English). In other words, is the speaker’s L1 translation equivalent (e.g., “berg” (mountain)) activated when the word “mountain” is being processed? They asked fluent Dutch-English bilinguals to name pictures in their L2 (English) and the distracter words (L2 in Experiment 1 and L1 in Experiment 2) were presented auditorily. The stimulus onset asynchrony (SOA) was manipulated in order to investigate the time course of the bilingual word processing.

The experimental conditions are summarized in Table 1. In Experiment 1, given a pictured mountain (“berg” in Dutch) for example, four types of L2 distracters were used: (1) phonologically related word (e.g., “mouth”), (2) phonologically related word to the Dutch translation equivalent (“bench” for the Dutch word “berg”) or “phono-Dutch,” (3) semantically related word (e.g., “valley”), and (4) unrelated distracters (e.g., “present”). In Experiment 2, given a pictured mountain, four types of L1 distracters were used: (1) phonologically related word (e.g., “mouw” (sleeve)), (2) phonologically
related word to the Dutch translation equivalent ("berm" (verge)) or "phono-
Dutch," (3) semantically related word (e.g., "dal" (valley)), and (4) unrelated
distracters (e.g., "kaars" (candle)). The logic of the experiments was that if
the non-selected Dutch word was activated during lexical processing, the pre-
sentation of phono-Dutch word (both in English and Dutch) would interfere
with the naming latencies. They also varied the SOAs in order to identify the
time course of lexical processing (e.g., lemma and lexeme). Semantic and
phonological distracter stimuli were also used to localize the processing stage
within the SOAs.

The results showed that the non-selected Dutch word was activated during
lemma selection but not during lexeme selection. It was also shown that not
only the non-selected word itself, but semantically related L1 words were
also activated in that the reaction times increased when the semantically relat-
ed L1 word was presented during the lemma processing. The latter finding
further supported the language non-specific activation view (see Costa, 2005,
for a review). Thus, lemmas of both intended and non-intended languages
receive activation from the conceptual level. However, whether or not phono-
logical information of both languages is also activated or is less clear (see
e.g., Costa, 2005, for reviews). What is clear is that intention to speak one
language cannot prevent the activation flow of the other language. This inter-
pretation is also supported by cross-linguistic word intrusion errors such as
blends as substitutions reported by Poulisse and Bongaerts (1994).

Language specific versus language nonspecific selection

If activation flow is language nonspecific, the issue of selection concerns
how bilinguals select words in the intended language. Whereas researchers
agree with the language non-specific activation view, there remain controver-
sies regarding how lexical selection takes place. There are two theoretical
views regarding this issue; namely, language specific versus language non-
specific selection views.

Language nonspecific and language specific selection models differ with
respect to whether the selection mechanism is affected by the semantically
activated lexical items of the unintended language. More specifically, lan-
guage specific selection models assume that the selection mechanism is
"blind" to, or ignorant of, the activation level of lexical items of the unintend-
ed language. In this sense, bilingual lexical selection is considered as similar
to monolingual lexical selection in that all competitors are activated lexical
items of the intended language (Costa, 2005). As Figure 1 shows, language
specific selection models predict that only within-language lexical items
(e.g., "CAT") are potential competitors in naming the pictured dog, and the
Spanish translation equivalent (i.e., "PERRO") and related lexical items (e.g.,
"GATO") have no impact on selection processes. Thus, however high activa-
tion levels of lexical items of the unintended language are, the selection
process is protected from activation levels of the unintended language items.
In contrast, language non-specific models assume that activated lexical items
of both the intended and unintended languages lead to competitions (see
Figure 1). Therefore, selection takes place by considering activation levels of
lexical items of both languages (e.g., "CAT," "PERRO," "GATO").
Figure 1. Models of the bilingual lexical selection system (based on Costa, 2005). The squares represent Spanish lexical nodes and the circles English lexical nodes. The arrows represent the activation flow and thickness of letters of the lexical nodes indicates activation levels. Part (a) represents language specific and part (b) represents language nonspecific selection mechanism.

Simple access, complex selection versus complex access, simple selection models

This dimension of “simple access, complex selection” versus “complex
access, simple selection” (La Heij, 2005) concerns specific implementations of lexical selection mechanisms. In a nutshell, this distinction concerns whether bilingual lexical selection is sufficient or not on the basis of conceptual activation only. Thus, “simple access, complex selection” models assume that access is simple (e.g., activation of lexical concepts) but the selection processes require sophisticated or intelligent mechanisms such as tags (Roelofs, 1992), binding-by-checking mechanisms (Roelofs, 1998) or inhibition mechanisms (Green, 1998a, b). In contrast, complex access, simple selection models assume that the conceptual input (which is the output of conceptual activation and selection) includes all necessary information (e.g., pragmatic information) for selection of the intended lexical item, so that the selection mechanism is grounded in on activation levels only. In this view, all the selection conditions (including words’ pragmatic and affective features; therefore, there is no convergence problem, according to La Heij, 2005) are prepared at the conceptual level (i.e., proactive) and the lexical selection is simply the matter of activation. Because the basic assumption of complex access, simple selection models is modularity (La Heij, 2005), no external information other than the conceptual input is required to select the right lexical item.

**Bilingual speech production models**

In this section, major L2 speech production models proposed to date will be reviewed. Before reviewing the models, it is useful to apply the two dimensions that I have already described in the previous section to the to-be-reviewed L2 production models. It should be noted, however, that the two dimensions are applicable to conceptual and lemma levels although the notion of language specific versus language non-specific selection can be
Table 2  Types of L2 Speech Production Models in Terms of the Manner of Lexical Access

<table>
<thead>
<tr>
<th></th>
<th>Simple access, complex selection</th>
<th>Complex access, simple selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language nonspecific</td>
<td>De Bot (1992, 2000)</td>
<td></td>
</tr>
</tbody>
</table>

also applicable to the phonological encoding stage (Costa, 2005). The following L2 production models are to be presented in a chronological order: De Bot (1992, 2000); De Bot and Schreuder (1993); Poulisse and Bongaerts (1994); Green (1998a, b); Kormos (2006). Table 2 presents the overview of the L2 speech production models in terms of the two dimensions; thus, the language specific versus language non-specific selection dimension and the simple access, complex selection versus complex access, simple selection dimension. With respect to Costa (2005) and La Heij (2005), because basic ideas of their speech production model are already described above, I will not repeat the descriptions of their models.


With respect to De Bot’s model, as in Lavelt, conceptualization is the first processing step, whose output is the preverbal message. It includes two processing steps; thus, macroplanning and microplanning. The former concerns the speaker’s choice of communicative goals and their elaborations and the latter concerns language specific pre-linguistic specifications as well as perspective taking. Following the conception of Levelt (1989), De Bot (1992, 2000) considers microplanning as language specific but macroplanning as non-language specific. The code-switching literature shows that motivations
for “intentional” code-switching are considered as reflections of the speaker’s intentions that reflect psychological, communicative, discoursal, situational and social factors (e.g., Dörnyei & Kormos, 1998; Hoffmann, 1991; Myers-Scotton, 2006; Sridhar, 1996). Accordingly, De Bot proposes that the information regarding which language to use is encoded during the microplanning and specified in the preverbal message. De Bot also assumes that the bilingual like the monolingual refers to the non-language-specific declarative knowledge.

Regarding the lexicon, De Bot adopts Paradis’ Subset Hypothesis that assumes that the bilingual lexicon has a network system. De Bot also adopts Levelt’s distinction between the lemma and the lexeme. He claims that the argument regarding whether the bilingual lexicon is one or two is oversimplified, pointing out that (1) a single storage of the lemmas is possible only when two languages have exactly similar in meanings and syntactic properties, and (2) there are cases where two languages have different lemmas (e.g., gender) while their forms are quite similar (e.g., some cognates and ambiguous words). Therefore, one-to-one correspondences between the lemmas and lexemes are difficult to retain. He then adopts Paradis’ Subset Hypothesis (1987), which assumes a single storage system and states that “in general, elements from one language will be more strongly connected to each other than to elements from another language, which results in the formation of subsets which appear to consist of elements from the same language, and which can be retrieved separately” (De Bot, 2000, p. 430). De Bot points out the compatibility of the Subset Hypothesis with the spreading-activation view of the mental lexicon and points out that the adoption of the spreading activation model makes the classic question of whether there are one or two language systems off the subject.

Regarding the lexical selection and the formulator, De Bot adopts Green’s
(1986) three-way distinction of selected, active, and dormant languages. According to Green (1986), languages within an individual can be classified into three types based on their activation states: selected language, which is selected in the ongoing speech; active language, which is not selected in the speech but can affect the ongoing speech (e.g., processing delay, interference, and also fluent code-switching); dormant language, which is in the long-term memory but has no influences on the ongoing speech. De Bot (1992, 2000) also adopts Green’s (1986) idea that both the selected and the active languages behave identically within the proposed model (i.e., the assumption of parallel formulators and their products of speech plans) except that the phonetic plan of the active language is also processed in the articulator. This means that De Bot postulates the single lexicon and the parallel processing of the two speech plans performed by the two language-specific formulators although the language-specificity of the formulators depends on the typology of the languages and the proficiency level.

Regarding the articulator, De Bot (1992, 2000) assumes that there is only one articulator that accesses a single large pool of L1 and L2 sounds and pitch patterns. For DeBot, this explains the fact that even the advanced bilingual speaker shows cross-linguistic influences at the phonological and pronunciation levels.

De Bot and Schreuder (1993)

In a subsequent paper, De Bot and Schreuder (1993) make several modifications to De Bot (1992) and propose a bilingual speech production model. First, they abandon the formulation of parallel speech plans and the notion of multiple formulators. Information regarding the lexicon and syntactic procedures from different language are assumed to form subsets and they are seen as a spreading activation network.
Second, the authors introduce Bierwisch and Schreuder’s (1992) notion of the verbalizer (Vbl), which is located between the conceptualizer and the formulator. According to the authors, the Vbl is necessary because the conceptualizer produces the preverbal message that may exceed the size of a single lemma and there are no one-to-one correspondences between concept primitives and semantic information; thus, the Vbl’s role is to curve up the conceptual structure so that concept-lemma mappings are optimized.

Another important modification to De Bot (1992) is that they propose that the preverbal message itself is not language specific though it contains language cues. The processing component that is in charge of language-specific lexicalization patterns here is the Vbl; thus, “[t]he Vbl function must know which conceptual primitives go together for a particular language and should curve up the conceptual structure accordingly” (p. 195).

De Bot and Schreuder (1993) also adopt a lemma-selection principle from Bierwisch and Schreuder (1992). Thus, for fast and accurate lemma selection, De Bot and Schreuder (1993) adopt Bierwisch and Schreuder’s (1992) matching principle, which states that “[a] semantic form triggers Lemma (i) if and only if there exists a complete match of all structures in the semantic form with all structures in the semantic representation of Lemma (i)” (p. 196).

Based on these assumptions, their bilingual production model is presented. First, De Bot and Schreuder (1993) assume the existence of a monitoring system, which monitors the discourse model for information regarding the conversational setting including the information regarding languages to be used. This information is subsequently delivered to the conceptualizer.

Second, according to their model, the preverbal message itself is not language-specific, but it carries information regarding which language to be used in the form of language cues. The values of language cues can vary
depending on the conversational setting.

Third, the Vbl breaks the preverbal message into language-specific chunks that are manageable to meet language-specific requirements for lexicalization patterns. During this chunking process, the Vbl considers the availability of specific items because “the Vbl ‘knows’ just as much as the rest of the Formulator” (p. 204). Both conceptual chunks and lexical items are also considered as having language cues, with which the Vbl tries to match the chunked information.

In addition, De Bot and Schreuder (1993) adopt a spreading activation mechanism based on the ideas of Paradis (1981) and Green (1986). Therefore, when a word from a language is activated, other words from the same language increase their activation levels. In the case of regular code switchers, words from different language subsets are assumed to have more or less strong connections as in word connections within a single subset; consequently, lexical items from different subsets have more or less equal activation, leading to random code switching.

Regarding language control, similarly to Paradis (1981) and Green (1986), they assume the existence of activation and deactivation mechanisms; however, this function has a certain limit. For instance, habitually used L2 items may have higher activation levels and are therefore difficult to suppress as in the situation where L2 words slip into L1 speech even when the intended word belongs to the L1 subset.

Fourth, with respect to formulating, similarly to Levelt (1989), relevant syntactic information of a selected language is assumed to become available after lemma selection. They assume that inventories of both morpho-syntactic rules or procedures and lexemes form subsets. Therefore, it is possible that activation spreads via similar form characteristics (e.g., phonemes and other phonological features). If that is the case, an L1 lemma spreads its activation
to the corresponding L1 lexeme, which in turn spreads its activation to an L2 lexeme that shares form information (e.g., cognates), leading to unintentional code switching.

According to Poulisse (1997), De Bot and Schreuder’s (1993) speech production model suffers from several problems. First, they claim that in the case of immigrants, random code switching is achieved by ignoring language cues because both language cues are equally important and lexical items from different languages are more or less equally activated. Or when the speaker tries to find an L1 word in time but when the value of the language cue is low, the speaker accepts an L2 words again by ignoring the language cue. Additionally, if there is no L2 item that accurately reflects the speaker’s intention, L2 speakers often rely on a “second best solution” (e.g., the use of compensatory strategies or L1), resulting in the selection of lexical elements that do not exactly reflect the original semantic specifications. In this case, an L1 word may be used in replacement of the intended L2 words. Clearly, these explanations violate their matching principle. As Poulisse points out, if the language cue can be ignored, the matching principle is no longer true.

Second, their model has a chunking problem. Poulisse (1997) refers to the acknowledgement by Bierwisch and Schreuder (1992) that exact ways to chunk conceptual primitives are not clarified. Poulisse then gives comments that if that is the case, chunking would be easier if the conceptualizer knows what lexicalization patterns are possible in a given language beforehand. According to Poulisse, however, the problem with this solution is that both Levelt (1989) and Bierwisch and Schreuder (1992) do not assume the conceptualizer to be accessible to the mental lexicon. In other words, the conceptualizer is “blind” in this regard. It may be possible to assume a feedback loop to the Vbl when the system finds out that the intended language does not allow certain lexicalization patterns; however, to do so violates Levelt’s
Poulisse and Bongaerts (1994)

Poulisse and Bongaerts (1994) is a study that aimed to develop a model of bilingual speech production based on analyses of non-intentional language switches. The data are from part of the Nijmegen project on the use of compensatory strategies (Poulisse, Bongaerts, & Kellerman, 1984; Poulisse, 1990). The language switches data were elicited from three different proficiency groups of Dutch learners of English, whose English learning experiences were three, five, and eight years respectively. The corpus of language switches contained 771 instances of unintentional L1 use during English speeches elicited by various tasks: oral description tasks of objects that the participants do not know their names in English (task I) and that have no conventional names (tasks II), four L2 story re-telling tasks based on short stories initially delivered in their L1 (task III), and a fifteen-minute talk about everyday topics with a native speaker of English (task IV).

It is well-documented that L2 learners often use intentional code-switching as communication strategies (e.g., Dörnyei & Kormos, 1998). Likewise, L2 learners often suffer unintentional L1 use during L2 speech (i.e., use of L1 content and function words as well as editing terms or fillers). Such instances of unintentional code-switching were ascribed by Giesbers (1989, cited in de Bot, 1992, 2000; Poulisse & Bongaerts, 1994) to language interference rather than to contextual and situational factors. In this sense, they were considered as slips of the tongue or speech errors. If that is the case, they should provide rich opportunities to investigate the mechanism of the bilingual speech production model as L1 speech production models were developed out of such inquiries. Especially, such data should provide the mechanism of how bilin-
goals keep the two languages separate and the explanation of why and how code-switching takes place.

In analyzing the data, Poulisse and Bongaerts propose their bilingual speech production model. As I reviewed above, De Bot (1992, 2000) adopts Paradis's (1987) Subset Hypothesis and claims that there is only one lexical store both for L1 and L2 lexical items, which is represented in the form of a network. De Bot (1992) also claims that the bilingual can process two speech plans in parallel. Regarding De Bot's proposals, Poulisse and Bongaerts (1994) point out two problems. The first problem concerns the effectuation of code-switching. The problem is that De Bot identifies the pre-linguistic encoding regarding the decision of language switch with micro-planning one the one hand and he assumes the parallel processing of the two speech plans on the other. The apparent problem here is that if the decision regarding language selection is made during the micro-planning, it is not clear why he postulates the parallel processing of the two speech plans during formulation. The other problem is that if the parallel processing is assumed, the number of speech plans should be unlimited. In other words, De Bot's claims seem to contradict with each other and it is not clear how bilingual lexical access is to take place.

According to Poulisse and Bongaerts (1994), an alternative and more economical explanation of separating the two languages is to assume language tags to be conceptual features or cues assigned during microplanning (see Figure 2). As Figure 2 shows, Poulisse and Bongaerts (1994) assumes that the conceptual level includes feature nodes with conceptual feature (e.g., [+ human]) and language nodes (e.g., [+ English]). Figure 2 also shows that similarly to De Bot (1992, 2000), Poulisse and Bongaerts (1994) assumes the decision regarding the language choice to be made during microplanning and the mechanism of bilingual lexical access to be based on the notion of
spreading activation. Poulisse and Bongaerts’ model predicts that if the speaker intends to produce the English word “boy,” the conceptual node with features of [+ human], [+ male], [- adult], and the language node of [+ English] are first activated and the activation spreads to the linked lemmas; subsequently, lemmas that share those conceptual features get activated in varying degrees (e.g., “JUGEN [+ Dutch],” “MAN [+ English],” “GIRL [+ English]” etc.). In this case, lemmas of the non-intended language (i.e., Dutch) of the bilingual also receive activation (i.e., non-selective access, see Kroll & Dijkstra, 2002; Kroll & Sunderman, 2003, for reviews). Among them, the one that receives the most activation is to be selected (i.e., the lemma “BOY [+ English]” in Figure 2).

The adoption of the spreading activation mechanism produces specific hypotheses. Poulisse and Bongaerts’ model predicts that if the spreading activation account of the bilingual speech production model is tenable, it should show frequency effects in the case of unintentional code-switching. For instance, Giesberts (1989, cited in De Bot (1992, 2000) and Poulisse & Bongaerts (1994)) reports that unintentional code-switching involves more
function words than content words and he explains the phenomena in two terms: (1) the speaker pays more attention to meaning than to form, and additionally (2) function words tend to be more frequent and their availability is higher than content words’ availability. Based on his claims, Poulisse and Bongaerts further hypothesized that if the bilingual speaker’s lexical access was based on the spread activation, it would also reveal proficiency effects as well (i.e., beginning > middle > advanced learners, where the symbol “>” means “is more frequent in the use of both L1 content and function words as well as editing terms during L2 speech than”) due to the difference in the amount of exposure to the L2.

Another issue Poulisse and Bongaerts raised concerns Myers-Scotton’s (1992) findings about the bilingual’s intra-sentential code-switching (more specifically, intra-clausal code-switching). Myers-Scotton (1992) makes two distinctions: one, following Levelt (1989), between the lemma and the lexeme, and the other, following Joshi (1985), between the Matrix Language (ML) (i.e., frame-building language) and the Embedded Language (EL). According to her, in intra-sentential code-switching, content morphemes from the Embedded Language are inserted into the morpho-syntactic frames of the Matrix Language. Myers-Scotton (1992) claim that “the bilingual speaker always accesses ML lemmas and builds the morpho-syntactic frames on the basis of the relevant information contained these lemmas” (Poulisse &

---

1 According to Myers-Scotton (2006), content morphemes are elements that either give or receive thematic roles (i.e., verbs and nouns). In addition, discourse markers (e.g., therefore, but, so) are also classified as content morphemes because they are considered as assigning thematic roles at the discourse level. System morphemes, on the other hand, are elements that do not assign or receive thematic roles. The classical distinction between content and function words does not necessarily correspond to the distinction. For example, pronouns are considered as function words, yet they receive thematic roles; therefore, they are classified as content morphemes in her model (see Myers-Scotton, 2006, pp. 244–256, pp. 267–271 for more details).
Bongaerts, 1994, p. 39) (see Myers-Scotton, 2006, for a review). As such, in intra-sentential code-switching, the morpho-syntactic frames are always given by the Matrix Language. Based on her theory, Myers-Scotton (1992) formulated two specific principles of the unequal participation of languages in intra-sentential code-switching: the morpheme order principle and the system morpheme principle. According to her, these principles are applicable to what she calls “mixed constituents,” which are “constituents including morphemes from both languages” and “[t]hese constituents may consist of an entire clause, but smaller phrases within the larger clause are also called mixed constituents” (Myers-Scotton, 2006, p. 244). The “morpheme order principle” states that in mixed constituents, the morpheme order is that of the Matrix Language and the “system morpheme principle” states that inflections and certain function words that stand alone come from the Matrix Language. Poulisse and Bongaerts (1994) tested whether these principles were also applicable to L2 speech, where the L2 served as the Matrix Language and the L1 served as the Embedded Language. They also attempted to offer some accounts regarding bilingual lexical access and morpho-phonological encoding processes during intra-sentential code-switching.

In order to analyze the data, Poulisse and Bongaerts (1994) classified the instances of unintentional lexical switches2 into [± morpho-phonologically adapted in the L2] and [+ content or function word]. Thus, four categories were made. In addition, L1 editing terms were coded separately (e.g., Dutch words “nee” (no) or “of” (or)) and were included in the [− morpho-phonologically adapted in the L2]. It should also be noted that many of the function

---

2 Poulisse and Bongaerts' (1994) definition of unintentional switches was based on hesitation phenomena and intonation. Thus, "we considered those cases to be unintentional which were not preceded by any signs of hesitation and did not stand out from the rest of the utterance by a marked intonation. When in doubt, we did not include the item in our analysis" (p. 43).
words belong to Myers-Scotton’s system morphemes. With respect to the effect of frequency and proficiency on the morpho-phonologically non-adapted unintentional switches (749 out of 771 instances), they found clear patterns (1) that their participants used more function words than content words in general (316 vs. 131 instances) and (2) that there were proficiency-related effects; namely, the frequency of the unintentional use of the L1 content and function words as well as the editing terms decreased as proficiency increased. Additionally, they examined self-repairs of L1 content and function words and found (3) that content words were more repaired than function words (about 54% vs. 31%) and content words were repaired more incomplete than function words (about 79% vs. 16%). Regarding the use of L1 words that were morpho-phonologically adapted, they found that such cases were rare (only 22 cases out of 771 unintentional switches). They then examined the cases based separately on morphological and phonological adaptations. There were 12 cases of morphological adaptation\(^3\) and 10 cases of phonological adaptations\(^4\) although these were not statistically tested due to the small number of instances. Regarding the former finding, they found (4) that all the morphological adaptations (i.e., inflections) except for one possible case were made with the English zero morpheme\(^5\), which the researchers interpreted as supporting Myers-Scotton’s claim that inflections in mixed

\(^3\) A sample switch was “you neem,” (“you take” in English) which was the combination of a Dutch base form plus English zero morpheme.

\(^4\) A sample switch was “stuk,” which is a translation equivalent of English word of “piece” and pronounced as “stuck” using an English phoneme /A/.

\(^5\) Poulisse and Bongaerts (1994) report that there were no instances of both Dutch verb base forms attached with Dutch inflectional morphemes and English base forms attached with Dutch inflectional morphemes. All Dutch words were thus combined with English zero morphemes with no instances of Dutch verb stems combined with other English inflectional morphemes (e.g., -s, -ing etc.) even when they appeared in contexts where English inflectional morphemes were possible. Note that these phenomena were limited to unintentional language switching not intentional language switching.
constituents come from only the Matrix Language (i.e., English in this case). There was also a finding (5) that showed the marked absence of morphologically adapted switches (zero case) and phonologically adapted switches (two cases) in the case of L1 function words (total 10 instances including content word switches). Because English function words do not require inflections, the former finding was not surprising. On the other hand, because there were only two instances of phonological adaptation of L1 function words to the L2, they concluded that “L1 function words, when they are used unintentionally in the L2, are hardly ever adapted to the L2” (p. 46).

What are the implications of those findings for models of L2 speech production? Regarding the findings (1) the greater use of L1 function words than content words and (2) the proficiency-related effect on the use of content and function words as well as editing terms, Poulisse and Bongaerts (1994) claim that these results supported the spreading activation account of the bilingual speech production model. They claim that because the amount of activation needed to access L1 function and content words is in general smaller than corresponding L2 lexical items, there was a higher probability that L1 words, especially L1 function words, would be erroneously selected, and these effects were reduced as proficiency increased due to the increased contact with the L2. They also claim that the effects are enhanced when the L2 learner has little attention, which is often associated with the beginning L2 learner. The latter claim is also related to both the higher use of L1 editing terms, where “the production of repairs will make extra processing demands on the speaker” (p. 47) and the finding (3) that content words were corrected before they were completed more than function words. According to the authors, the reason is that because the speaker and especially “beginning learners generally do not have much attention to spare, they may need to make a choice between the items to which they are going to attend” (p. 47). By assuming the
spreading activation account of lemma access, Poulisse and Bongaerts (1994) claim that errors of selecting L1 lemmas for the intended L2 lemmas are explained in a similar way as in L1 substitution errors (e.g., “high” for “low”).

Regarding the finding (4) that all the inflectional adaptations were made with the English zero morpheme, which was taken to support Myers-Scotton’s claim that inflections in mixed constituents come from only the Matrix Language, Poulisse and Bongaerts (1994) have further advanced the argument. They referred to Levelt’s (1989) account of morphological encoding, which specifies that inflectional word forms have separate form representations in the mental lexicon. They claim that if the word forms are represented as Levelt assumes in the L1 speech production, there should be such instances as “you neemt” (second-person singular form) because in this case a Dutch verb lemma is accessed, which has Dutch diacritic parameters; however, they found otherwise. In view of this, the authors claim that there should be another way for L2 learners to produce words. Specifically, they claim that both verb stems and inflectional morphemes (e.g., -ed, -s, -ing) are accessed separately and are combined into single verb forms. More specifically, they claim that both verb stems and inflectional morphemes have their own lemmas, which are activated based on the semantic information in the preverbal message (e.g., tense, number etc.). Thus “[a]fter the lemmas for the

6 Although the authors claim that their results support Myers-Scotton’s hypothesis, given the presence of Dutch verbs with English zero morphemes only and the absence of Dutch verbs combined with other English inflectional morphemes, the results themselves seem to provide partial support. However, it should be noted that because their participants were non-balanced L2 speakers it is still possible to consider them as moving towards the end of the continuum (i.e., balanced bilinguals).

7 For instance, the word forms of “gives” and “gave” can be expressed respectively as [give-lemma; singular-number; third-person; present-tense] and [give-lemma; singular-number; second-person; past-tense].
base form and the inflectional morphemes have been accessed, the information contained in these lemmas can be used to set up a Levelt-like address frame pointing at one word form” (Poulisse & Bongaerts, 1994, p. 50). Poulisse and Bongaerts (1994) claim that the observed L1 switch errors such as “you neem” can be explained as a result of the erroneous access to the L1 base form and the correct access to the L2 morpheme (i.e., zero morpheme). According to them, such decomposed lemma representations for word forms can also account for speech errors such as “teached” or “gaved” made by L1 children and L2 learners. They also point out possible problems of this decomposed lexicon account and further assume that irregular verbs such as “gave” have their own full inflected lemmas and there is a checking device that intercepts forms that are not present in the lemmas. According to them, because the checking device does fail to intercept forms that do not exist in the lemma, they observed only a single instance whose inflection was adapted to the L1 (“one is won(ing)”) in their database. In short, Poulisse and Bongaerts (1994) claim that the lemma level contains both fully inflected lemmas for irregular verbs and decomposed lemmas for regular verb stems and their inflections, all of which have their language tags.

Finally, regarding the phonologically adapted switches, Poulisse and Bongaerts (1994) found there were only 10 phonologically adapted switches while they found a much larger number of non-adapted switches (749 instances). With respect to the 10 cases (e.g., “stuck” for the Dutch “stuk”), learners pronounced majority of Dutch words as if they were English words by replacing a Dutch phoneme with an English one. In view of these findings, they concluded that L2 learners normally encode L2 lemmas by applying L2 phonological encoding and encode L1 lemmas by applying L1 phonological encoding, yet, the presence of the phonologically non-adapted switches support De Bot’s (1992) claim that there is one articulator that draws on
one large pool of sounds and pitch patterns tagged for language, which sometimes erroneously encodes L1 lemmas with L2 phonemes. Accordingly, phonologically non-adapted switches are explained as erroneous access to the L1 lemma (e.g., *stuk*), which is subsequently processed by L2 phonological encoder ("stuck" using an English phoneme /n/).

Poulisse and Bongaerts' (1994) L2 speech production model can be summarized as follows. First, regarding conceptualization, the information regarding the decision to select the intended language is coded during microplanning and represented in the preverbal message as De Bot (1992, 2000) claims; however, unlike De Bot's proposal, language-feature nodes are postulated at the conceptual level. Thus, language cues control the quality of the input to the lemma level. Second, regarding the lemma selection, the selection mechanism is implemented on the basis of the spreading activation only and each lemma is tagged for language⁸ (i.e., simple selection in La Heij’s sense). This model assumes that lemmas from both languages receive activation (i.e., language non-specific activation model, see e.g., Costa, 2005; Hermans, Bongaerts, De Bot, & Schreuder, 1998; Kroll & Dijkstra, 2002; Kroll & Sunderman, 2003) and in selecting the intended lemma, activation levels of lemmas of the non-intended language is also taken into account (i.e., language non-specific selection, see Costa, 2005). As such, errors due to erroneous selection of L1 lemmas for the intended L2 lemmas are treated similarly to semantically conditioned L1 substitution errors. The level of lemma

---

⁸ Regarding language tags, as La Heij (2005) mentions, Poulisse and Bongaert (1994) consider these tags as part of semantic information of lemma within the original framework presented by Levelt (1989). Whether lemmas include semantic information or not is not decisive; however, if the lemma does not contain semantic information, as La Heij (2005) points out, the presence of language cues at the conceptual level suffices at least within Poulisse and Bongaert's framework, which is the only speech production model including L1 and L2 speech production models that meets theoretical conditions of the complex access simple selection model proposed by La Heij (2005).
activation is sensitive to relative frequency and proficiency of the L2 speaker. Furthermore, irregular verb lemmas are represented as fully inflected lemma nodes, which are accessed directly, whereas regular verb lemmas are represented as decomposed lemmas; namely, verb stem lemmas and inflectional lemmas are represented as separate nodes at the lemma level. As a consequence, inflectional lemmas are conceptually conditioned (e.g., tense, number, person etc.). Third, regarding morphological encoding, the aforementioned verb stem lemmas and inflectional lemmas are assumed to form respective morphological address. Furthermore, inflection morphemes come from the Matrix Language, which is in line with Myers-Scotton’s claim. Forth, regarding phonological encoding and articulation, similarly to De Bot (1992, 2000), there is a single pool of L1 and L2 sounds and pitch patterns tagged for language. Normally, phonological encoding is determined by the language of the selected lemma whereas L1 words are sometimes erroneously encoded with L2 phonemes, which is sent further to the common articulator. Phonologically adapted L1 switch errors (e.g., “stuck” for “stuk”) are explained in terms of erroneous selection of L1 lemmas, which are then applied L2 phonological encoding procedures. Fifth and finally, they claim that there is a monitor that checks the combination of regular verb lemmas and their inflection morphemes against the available word forms once they are phonologically coded; if there is no match, the checking device intercepts the processing, but this leads to occasional detection errors. As such, their model can account for the separation and mixing two languages, fast code-switching without using active inhibition mechanisms (cf. Green, 1986a, b, 1998) but with language cues, and error sources (e.g., priming effect, incorrect assignment of language cues etc.) (La Heij, 2005). As La Heij (2005) points out, Poulisse and Bongaert’s speech production model is characterized by “complex access (to lemmas)” in the sense that conceptual features includ-
ing language cues need to be specified for lemma selection and by "simple selection" in the sense that lemma selection is determined by the activation level only. Thus, regarding bilingual lexical access, the model is characterized by three design features: (1) non-language specific activation, (2) non-language specific selection, and (3) complex access and simple selection.

Green (1998a, b)

In his inhibitory control (IC) model of bilingual language control through biasing activation levels of response, Green (1986, 1993, 1998a, b; also see Kroll & Dijkstra, 2002) adopts a model of action control proposed by Norman and Shallice (1986), often referred to as Norman-Shallice model (also see Monsell, 1996). In this section, I will first describe the notion of action schema and afterwards Green’s bilingual production model.

The Norman-Shallice model assumes multiple levels of action control and Green (1998a, b) adopts their model. According to Norman and Shallice (1986, also see Rumelhart & Norman, 1982; Shallice, 1994), action sequences are controlled by action or thought schemas. Schemas are "organized according to the particular action sequences of which they are a part, awaiting the appropriate set of conditions so that they can become selected to control action" (Norman & Shallice, 1986, p. 1; also see De Groot, 1998, for a review of the notion of “schema”). Thus, there can be higher-order schemas called “source schemas,” which call “component schemas” or sub-routines (Rumelhart & Norman, 1982), so that they can be hierarchically organized. In this regard, schemas are analogous to Kempen and Hoenkamp’s (1987) procedure call hierarchy, where higher-order syntactic procedures call their sub-procedures or sub-routines.

In performing certain habitual actions, particular schemas need to be selected in a right timing and coordinated (e.g., Monsell, 1996). Task
schemas, given external cues, often compete with each other in order to determine their activation levels, and the selection is assumed to be conducted directly by a regulating mechanism called "contention scheduling." Contention scheduling is a mutual conflict resolution process (Shallice, 1994). Thus, schemas activate and/or inhibit one another and the selection of a schema takes place when its activation value exceeds the threshold value (also see Anderson, 1983, for a similar assumption in ACT*). The activation level of schemas is also affected by prior activation or priming. Furthermore, timing of particular actions is controlled by "triggering conditions" (i.e., under what circumstances) that allow activated schemas to be initiated at a particular point of time. For instance, Rumelhart and Norman (1982) implemented a computerized model of skilled typists and the triggering conditions were given to "key press schemas" as feedback information regarding the current positions of the fingers. This is necessary because the schema needs to specify which finger to press the target key given the current finger positions. More specifically, each schema activates or inhibits each other, determining the order of particular actions. For instance, in typing the word "very," a word schema "very" activates its sub-schemas called "key press schemas." When activated, the key press schema "v" inhibits other schemas to follow (i.e., key press schemas for "e" "r" and "y") so that the temporal action sequence is determined. As such, the order of activation values reflects the order of actions. Once the key press action is effected, the key press schema decreases its activation by itself and the inhibition on the other key press schemas is released. Thus, the selection of a particular schema is determined by the activation level under specific conditions. The essential characteristic of the mechanism is that it is automatic, bottom-up processes or exogenous control system.

Activation levels of schemas, in contrast to bottom-up action control, can
be affected indirectly by the supervisory attentional system (SAS), which is located at the highest processing level. According to Norman and Shallice, the intervention of the SAS (i.e., deliberate attentional use) is required particularly when (1) complex planning or decision making is required, (2) troubleshooting or error correction is involved, (3) tasks are novel, and (4) the task is dangerous or technically difficult (Norman & Shallice, 1986). Thus, deliberate attentional control is required "for fear that it might lead to error" (Norman & Shallice, 1986, p. 3). The major advantage of Norman-Shallice model is that as Monsell (1996) and Styles (1997, 2005) point out, the multiple-level action control is more elaborated than the simple dichotomy of controlled versus automatic processing.

Adopting the basic assumption of the Norman-Shallice model, Green (1998a, b) proposes a bilingual speech production model (Figure 3). As Figure 3 shows, there are non-linguistic processing components (i.e., conceptualizer, the SAS, language task schemas) and a linguistic component (i.e., bilingual lexico-semantic system). Basically, Green adopts Levelt’s model, where activated non-decomposed lexical concepts spread their activation to the lemmas connected to them. What distinguishes his speech production model from others is the explicit formulation of bilingual production control system.

As mentioned above, Green adopts the notion of action schemas or in his terminology “task schemas.” Task schemas (e.g., L1-to-L2 translation schema, L2 word production schema) can be established by the SAS (or task set (re-)configuration in Monsell’s (1996, 2003) term), so that the system can achieve the task goal (represented as “G” in Figure 3). For example, presenting a word “dog” can elicit potentially several responses, such as producing an L2 translation equivalent, simply reading it, or producing its antonym and so forth. This means that the selection process is potentially competitive. In
In order to achieve the current task goal (e.g., orally translating an L1 word into an L2 one), L1 word production schema needs to be actively suppressed; thus, the L1-to-L2 translation schema calls and boosts L2 word production schema, which actively suppresses L1 word production schema. Such action control steps, according to Green (1998a, b), take place independently of linguistic or lexico-semantic processing (see Green, 1986, 1993, 1998b, for clinical evidence of brain-damaged patients). As such, established task schemas constitute what Green calls “functional control circuits” to achieve specific task goals that are produced by the conceptualizer and maintained by the SAS during task performance.

In order to exert control over the lexico-semantic system, there must be a locus of lexical selection. Green assumes that each lemma has its language tag for either L1 or L2 and the language tag is specified during language-independent conceptualization, which also affects but does not determine the
Figure 4. A simplified version of the Inhibitory Control Mechanism (Green, 1998) based on Finkbeiner, Gollan, and Caramazza (2006). In this example, the target language is English (the speaker’s L1), and the non-target language is Spanish (L2). The inhibitory connections between the language task schemas and L2 lexical nodes indicate suppression of lexical nodes in the non-target language.

Lexical selection. According to Green, the IC model assumes that the locus of lexical selection is the lemma level.

Regarding the lexical selection mechanism, the IC model employs several checking procedures and inhibition mechanisms. Thus, Green assumes that there is a language-independent conceptualizer and the mapping of thought onto the word is based on the mechanism of binding-by-checking between the lexical concept and its lemma (Levelt et al., 1999). In addition to regular lexical selection, in the case of translating from L2 to L1, according to Green, L2 lemmas point to the L1 lemmas of translation equivalents so that the binding-by-checking mechanism ensures that the correct L1 translation equiva-
lents are linked to the correct L2 lemmas. However, since the language tag is not a decisive factor in the IC model (also see Kroll, Bobb, & Wodniecka, 2008) and often incorrect lemmas can be activated if they share conceptual properties with activated lexical concepts, further inhibitory mechanisms are necessary that modulate lexical competition by suppressing lemmas with incorrect tags. For instance, when an English native speaker attempts to produce the word “dog” in translating its Japanese translation equivalent “inu,” the binding-by-checking mechanism ensures that the English lemma dog is linked to the activated L2 lemma inu.9 In this case, both inu and dog lemmas are activated, though the Japanese word must be inhibited or else it may capture speech production. Initially, L2-to-L1 translation schema is also activated and because the production needs to be done in the L2, the translation schema inhibits L1 word production schema and activates L2 word production schema. Language tags are utilized by the L2 word production schema by suppressing lemmas of the unintended L1 languages reactively as opposed to proactively. As implied by the postulation of these active inhibition processes (also see Lee & Williams, 2001), Green considers conceptual language specification as a necessary but not sufficient condition for selecting the intended lemma of a given language (also see Kroll, Bobb, & Wodniecka, 2008).

All in all, two collusions can be drawn. First, considering the supposition of the inhibitory mechanism, it is safe to say that Green’s model is classified as a simple access, complex selection model in La Heij’s sense. Furthermore, thanks to the tag suppression (e.g., exerted by L1 word production schemas),

9 Green (1998a, b) assumes that in the case of the forward translation (i.e., from L1 to L2), L1 word forms access lexical concepts via L1 lemmas. On the other hand, in the case of the backward translation, activation is assumed to spread via lexical links between L1 and L2 lemmas for the word association hypothesis (also see Kroll & Stewart, 1994; Kroll & De Groot, 1997).
competition is assumed to take place among lemmas within the intended language (i.e., lemmas not suppressed by word production schemas) and it is at this moment that the bilingual behaves like the monolingual in selecting the intended lemma. Although inhibition works, and if it does, reactively and there may be times when it does not have to do so, the IC model can be classified as a version of the language-specific selection model at least after the inhibition mechanism is engaged during the lexical selection.

*Kormos (2006)*

Kormos' (2006) model of L2 speech production is based on thorough reviews of L1 and L2 speech production literature and the newest one among the ones described in this article (Figure 5). Her L2 production model is based on Levelt's (1989) on the ground that his model is “the empirically best supported theory of monolingual speech processing” (Kormos, 2006, p. 166). Accordingly, her L2 speech production model is modular, consisting of the conceptualizer, formulator, and the articulator, based on which incremental parallel language processing can be realized if the learner's proficiency reaches a certain threshold\(^\text{10}\).

In Kormos' model, there is only one large declarative knowledge source or long-term memory, which itself contains different types of knowledge. This view contrasts with Levelt's in that his model assumes different knowledge sources: (discourse-internal and discourse-external) environmental knowledge stores, the lexicon, and the syllabary. Her model is based on

\(^{10}\) More specifically, Kormos (2006) assumes that, in contrast to Levelt's model, L2 *cascading* from lexical to phonological levels is possible. This means that L2 phonological nodes can be activated before lexical nodes are selected. In this sense, her model is not strictly serial as Levelt's. On the other hand, consistent with her modular view, backward activation from the phonological to the lexical levels is not assumed; activation can be leaked but never return.
Tulving's classic distinction between episodic and semantic memory, both of which are assumed to be subcomponents of long-term memory. Thus, episodic memory stores information about personal experiences (e.g., temporal and perceptual information) whereas semantic memory stores knowledge about facts and is assumed to be necessary for language use (Tulving, 1972); therefore, concepts, lemmas, lexemes, the syllabary, and L2 declarative knowledge are all assumed to be part of semantic memory that are shared in both L1 and L2. Along with linguistic concepts, semantic memory is also assumed to store non-linguistic concepts as well as meaning-related memory traces (Hintzman, 1986). In addition, L2 declarative rule knowledge is added to Levelt's model with recourse to Ullman's (2001) declarative and procedural
(DP) model.

L2 speaking starts with conceptualizing the message and deciding which language is to be chosen for speaking. During conceptualization, contextually relevant L1 and L2 concepts, stored together in semantic memory, are co-activated when they share conceptual features (i.e., conceptual activation). Subsequently, conceptual selection takes place (e.g., Bloem, van den Boogaard, & La Heij, 2004; La Heij, 2005) and only the intended concept in the intended language is chosen for subsequent processing. This conceptual selection means that not all activated concepts automatically activate their corresponding lexical items. In other words, conceptual selection means that the intended concept is selected as part of the preverbal message or for the process of lexicalization (Bloem, van den Boogaard, & La Heij, 2004). After conceptual selection, only the selected concept in the intended language spreads its activation to semantically related lemmas that include lemmas in the non-intended language (i.e., “semantic cohort activation,” Bloem et al., 2004). Returning to conceptualization, the extent of concept sharing between L1 and L2 is represented in the form of hierarchical conceptual networks and is dependent on types of concepts (e.g., concrete/abstract), the learning history on the individual (e.g., EFL, ESL), and the learner’s proficiency level. During conceptualization, it is also assumed that language cues are assigned (e.g., [+ English]) to each selected concept contained in the preverbal message so that flexible code-switching can be performed. This means that Kormos’ L2 speech production model does not assume complex selection mechanism. Finally, Kormos also assumes that routinely activated conceptual chunks associated with communicative functions, such as apologizing, become activated as single chunks. These conceptual chunks are assumed to activate the corresponding linguistic chunks (i.e., idioms, formulaic expressions), which are represented as single lemmas and retrieved as a whole.
With respect to lexical encoding, Kormos assumes that lemmas that are associated with the selected concept become activated (i.e., semantic cohort activation and language non-specific activation). Consequently, activated L1 and L2 lemmas enter into competition and the lemma that matches all the conceptual specifications gets selected. In contrast to Levelt, et al. (1999) and Green (1998a, b), Kormos’ model thus presupposes neither the checking mechanism nor the inhibition mechanism; her model can be classified as a “complex access, simple selection” model (e.g., La Heij, 2005; Poulisse & Bongaerts, 1994). She also speculates that L1 and L2 lemmas and lexemes are connected with each other both within and between languages and also that connection strength between lexical entries is assumed to vary and asymmetrical activation directionality (e.g., from L1 to L2 and vice versa) also needs to be considered (see Abutalebi & Green, 2008), with its scope including the notions of passive and active vocabulary (Meara, 1997).

After lemma selection, syntactic information in the lemma becomes activated and L2 syntactic encoding is launched. Kormos’ model assumes that similarly to Levelt (1989), L2 syntactic encoding follows the processing steps described by Kempen and Hoenkamp’s (1987) Incremental Procedural Grammar (IPG). Thus, her model of L2 syntactic encoding is also lexically driven. But she also acknowledges that research on L2 syntactic encoding is rather limited (see, Hartsuiker, Pickering, and Veltkamp, 2004; Meijer and Fox Tree, 2003, for studies on cross-linguistic syntactic priming). More studies are needed in this sub-domain and Kormos herself acknowledges that her proposal is in fact speculative given the limited number of experimental studies.

Kormos claims that unbalanced bilinguals, whose proficiency is low, frequently demonstrate characteristic linguistic behaviors. For instance, low proficiency learners are observed to commit L1 transfer errors. Such transfer
errors are due to differential cross-linguistic patterns of lexicalization (e.g., “enter into the room,” which Japanese and Hungarian learners of English often produce). This is considered as a result of associations between L2 lemmas and their corresponding L1 lemmas’ syntactic information available in the mental lexicon. When syntactic building procedures are concerned, those low proficiency learners are assumed to employ L2 rules consciously (i.e., non-automatic) or L1 rules are consciously applied to the L2 as communication strategies.

Similarly to other L2 speech processing mechanisms, Kormos also assumes that mechanisms for L2 phonological encoding, which involves the activation of the phonological form of the word to be encoded, syllabification, and setting the parameters for the loudness, pitch, and duration of international phrases, are basically the same as L1 speech production mechanisms. In contrast to the strictly discrete information processing flow assumed by Levelt et al. (1999), however, Kormos assumes that L1 and L2 lexemes are also become activated due to cascading though backward spreading activation is not assumed. This means that competitions for selection occur not only at the lemma level but also at the lexeme level. Following Roelofs (2003), L2 phonemes are also assumed to be stored as single units (e.g., [b], [p], etc.), not based on their phonological features (e.g., [+ voiced], [- nasal]), in a single network and identical phoneme representations between L1 and L2 are shared. For low proficiency learners, L2 phonemes tend to be equated with existing L1 phonemes, a process frequently associated with fossilization (Celce-Murcia, Brinton, & Goodwin, 1999). Those low proficiency L2 learners might need to resort to the declarative knowledge of phonological encoding or transfer L1 rules when the rule is not available.

With respect to phonetic encoding, Kormos’ view is similar to those of De Bot (1992) and Poulisse and Bongaerts (1994). Thus, L1 and L2 syllable pro-
grams are stored together in the single syllabary. Beginning L2 learners are assumed to use L1 syllable programs frequently whereas advanced L2 learners create and rely on L2 syllables.

Finally, with respect to monitoring, following Levelt's perceptual loop theory, three types of monitoring loops are assumed in L2 speech production. The first loop involves the comparison between the preverbal plan and the original communicative intention. The second monitoring loop, which is also called covert monitoring, involves the use of the internal loops (i.e., monitoring of "inner speech" or the phonetic plan). The third monitoring loop concerns the monitoring of overt speech that employs the external loop. Similarly to Levelt, Kormos assumes that monitoring involves the use of the speech comprehension system. She also claims that because L2 learners' attentional resource is limited and they need to conduct monitoring more frequently than the L1 counterparts, they have little attention available for monitoring.

Further issues in bilingual speech production

This article described some critical notions for classifying bilingual speech production models and representative bilingual speech production model. In this section, I will also refer to some other critical issues that are not addressed in the previous sections in three terms: locus of language selection and inhibitory mechanisms.

Whereas the previous sections described bilingual lexical selection in terms of manner of selection, it is also possible to describe and classify bilingual speech production models in terms of locus of language selection (e.g., Kroll, Bobb, & Wodniecka, 2008; Meuter 2005). Table 3 summarizes types of bilingual speech production models in terms of language selection. For
instance, a language selection model that assumes that language selection takes place at the conceptual level is considered as a "language selective." In this type of model, language selection usually specified by a language tag is powerful enough to determine which language the speaker will choose. All the speech production models that were described in the previous sections, except for Green's, belong to category in this regard. In Green's model, however, the language tag can affect but does not determine the language selection, which takes place at the lemma level; therefore, it is considered as a language nonselective model whose hypothesized competition is to take place at the lemma level. For more detailed descriptions and examples of each type, refer to Kroll, Bobb, and Wodniecka (2008).

Kroll, Bobb, and Wodniecka (2008), while arguing for nonselective models of bilingual speech production, do not support any of those bilingual speech production models that have fixed loci of language selection, but they claim for variable loci of language selection and point out several factors that

<table>
<thead>
<tr>
<th>Locus of language selection</th>
<th>Language selective</th>
<th>Language nonselective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determined at the conceptual level</td>
<td>Lemma/abstract lexical level</td>
<td>Beyond the phonological level</td>
</tr>
<tr>
<td>(a) following feedback from the phonology to the lemma level, or</td>
<td>Phonological level</td>
<td>(b) during the execution of the articulatory plan</td>
</tr>
<tr>
<td>(b) during the execution of the articulatory plan</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of hypothesized competition</th>
<th>Only candidates within the target language</th>
<th>Within- and across-language lexical candidates compete</th>
<th>Within- and across-language lexical and/or phonological candidates compete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within- and across-language lexical, phonological, and/or articulatory features compete</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Alternative Models of Bilingual Language Selection (based on Kroll, Bobb, & Wodniecka, 2008)
affect them. Those factors include: language proficiency, relative dominance of the speaker’s languages, task demands, degree of activation of non-target items and so on. This is an interesting proposal given the fact that any of the current bilingual speech production models cannot accommodate existing experimental data (e.g., Finkbeiner, Gollan, & Caramazza, 2006).

Another issue that needs to be mentioned is whether or not bilinguals use inhibitory mechanisms in achieving accurate language performance. Among the major bilingual speech production models addressed in this article, only Green assumes the inhibitory mechanism in his IC model. Although there are some objections to Green’s model (e.g., Finkbeier, Almeida, Jansen, & Caramazza, 2006; Finkbeiner, Gollan, & Caramazza, 2006), Abutalebi and Green (2008), drawing on functional neuroimaging research results, claim for the view that language switching is an example of task switching (e.g., Allport, Styles, & Hsieh, 1994; Monsell, 1996, 2003) and postulating multiple neural regions (see Monsell & Driver, 2000, for homunculus issues) of control that may rely upon an inhibitory mechanism is a necessary condition for the development of the bilingual speech production model. They further argue that prefrontal-basal-ganglia circuits, which they identify as the locus of inhibitory mechanisms, are engaged when bilinguals access a given language and that inhibitory mechanisms are key and bilingual production models in which inhibition plays no role should be ruled out through systematic examinations using neuro-imaging techniques.

**Conclusion**

This article first described some of the important notions concerning manner of lexical selection in the study of bilingual speech production. Afterwards, representative bilingual speech production models were
reviewed, followed by further reference to other important criteria that can help classify and examine various models of bilingual speech production: locus of selection and inhibitory mechanisms. Clearly, these classification criteria apply to bilingual speech production models independently and future studies should examine bilingual speech production models in those terms systematically.

References


Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 109–121). Cambridge, England: Cambridge University Press.


