Part VI

Control
Selection and control in bilingual comprehension and production

21.1 Introduction

Two observations about bilingualism have dominated the recent literature. One is the discovery that the bilingual’s two languages are continuously active, even when the task and context require that only one language be used. The co-activation of the bilingual’s two languages has been observed in the simplest of tasks, when naming words (e.g., Schwartz, Kroll, & Diaz, 2007), naming pictures (e.g., Costa et al., 2000), and recognizing spoken words (e.g., Marian & Spivey, 2003b). But it has also been shown to persist when these tasks are placed in sentence context (e.g., Libben & Titone, 2009) and when features of the two languages are markedly different, as in the case of two languages that utilize different written scripts (e.g., Hoshino & Kroll, 2008) or when one language is written and the other language is signed (e.g., Morford et al., 2011). The observation that it is not simple for adult bilinguals to exploit cross-language differences that might serve to separate the two languages makes the problem of language selection complex. It suggests that there is a high level of cross-language activation and potential competition when the intended language is selected. The fact that some groups of bilinguals code switch, changing language while in the midst of an utterance, provides additional support for the claim that the parallel activation of the two languages is not an obstacle, but a feature that can be exploited.

The second observation is that experience in using two or more languages has consequences not only for language processing itself but also
for cognition and the brain. Bilinguals, at all points across the lifespan, have been shown to differ from monolinguals in their ability to negotiate competition across alternative responses, to switch from one task to another, and to ignore irrelevant information (e.g., Bialystok, Craik, & Luk, 2012). Critically, the brain networks that enable cognitive control are different in bilinguals and monolinguals, with evidence that bilinguals are able to engage control mechanisms more efficiently than their monolingual peers (e.g., Abutalebi et al., 2012; Gold et al., 2013). What is notable in these studies of cognitive control is that the tasks that have been used to assess cognitive performance and brain activation are themselves not language tasks. Rather, they are non-verbal cognitive tasks that induce conflict (e.g., Stroop tasks, flanker tasks, the Simon task).

The initial interpretation of these two set of observations was that they are related. Bilinguals must negotiate the stream of competition across their two languages and experience in doing so creates expertise in the realm of cognitive control. Indeed, we have used the metaphor of the juggler to describe the situation that bilinguals face (e.g., Kroll et al., 2012). Learning to juggle two languages is hypothesized to produce skill in other cognitive domains in which similar types of competitive processes are manifest. But what aspect of linguistic athleticism produces these changes in cognition and the brain? Bilinguals do many different things with language and it is not at all clear that all or even most of them necessarily have any sort of direct correspondence with the observed cognitive and neural consequences.

Brain imaging studies show that the bilingual’s two languages share the same neural tissue (e.g., Abutalebi, Cappa, & Perani, 2005). When there are differences in patterns of brain activation for the first (L1) and second (L2) languages, they are likely to reflect the level of proficiency or skill associated with the two languages, with the L2 more likely to require additional cognitive control (e.g., Abutalebi & Green, 2007). The requirement to engage control mechanisms in using the L2 and in regulating the activity of the L1 seems likely to be the source of the observed consequences of bilingualism. At the same time, there is no accepted model to date of how language processes differentially recruit cognitive resources to enable fluent performance and to produce domain-general consequences.

In the recent literature, there are a set of proposals that begin to address the broader issues of how bilingual experience shapes the neural networks that support language and its cognitive consequences (e.g., Baum & Titone, 2014; Green & Abutalebi, 2013; Kroll & Bialystok, 2013). In the present chapter, we focus more narrowly on the processes that enable bilinguals to select the intended language when they read text, listen to spoken

---

1 There are a few recent papers that attempt to catch the cognitive consequences of language processing as it takes place online (e.g., Blumenfeld & Marian, 2011; Martin, Macizo, & Bajo, 2010; Wu & Thiery, 2013). We discuss this work later in the chapter.
language, or plan speech in one language or the other. We return at the end of our discussion to consider how mechanisms of selection, which, as we will show, may differ for comprehension and production, may hold a clue to better understanding the cognitive consequences of multiple language use. Our review and analysis considers evidence primarily from lexical comprehension and production because that is where there has been the greatest number of studies and where these processes can be compared systematically. There is a valid and important question to be asked about how the claims about lexical processing might be extended to the comprehension and production of phrases, sentences, and discourse, but that question is beyond the scope of the present chapter.

As we noted at the onset, there is compelling evidence in all domains that bilinguals activate both languages in comprehension and in production. Curiously, there has been little direct comparison of the evidence in comprehension and production. Cross-language co-activation may be resolved quite differently in comprehension than in production because the intended language may not always be explicitly selected in comprehension. In language production, selection is forced by the fact that, at least for unimodal bilinguals who speak each of the two languages, there is only one output channel, making it physically impossible for a bilingual to produce both languages at the same time. In contrast, in comprehension, it may not be a requirement to resolve co-activation fully. In this sense, language selection may not be necessary, or necessary to the same degree, during language comprehension. The goal of our chapter is to begin to compare these language processes as means to identify the scope of the control mechanisms that are engaged when bilinguals use each language alone or both languages together in mixed language contexts.

### 21.2 Bilingual comprehension

The evidence for parallel activation during language comprehension comes primarily from processing effects involving words with cross-language overlap, or words that are ambiguous between two languages, such as cognates and interlingual homographs (for a review, see Dijkstra, 2005). Cognates – translations which share both lexical form and meaning – often exhibit facilitated processing compared to lexically matched non-cognate control words due to the overlap at all levels (orthography, phonology, and semantics). In contrast, homographs are words that share lexical form but not meaning. To illustrate, the word “hotel” is a cognate in Dutch and English, with virtually the same form and meaning in both languages. In contrast, the word “room” means cream for your coffee in Dutch and is pronounced differently in the two languages. Homographs, also called false friends, often exhibit slower processing compared to
lexically matched controls, but depending on the context in which they are presented, they can exhibit facilitated processing (e.g., Dijkstra, Van Jaarsveld, & Ten Brinke, 1998; Lemhöfer & Dijkstra, 2004). Presumably, homographs facilitate processing at the levels of lexical form, but the conflicting semantics create competition that produces interference. The form of the resulting homograph effect will then depend on the extent to which task demands require different levels of processing. Theoretically, the facilitation observed in cognate and homograph processing is presumed to stem from heightened activation levels, a decreased threshold for activation, increased functional frequency, common lexical storage, or a combination of these factors (e.g., Dijkstra, 2005). Regardless of the exact locus of the facilitation, all accounts attribute the effects of cross-language interaction to the overlap of lexical codes. Cognate and homograph effects can occur only for speakers of two or more languages and not for monolingual speakers of any language alone; they are considered the hallmark of language co-activation or non-selectivity. Models like the Bilingual Interaction Activation Model (BIA+) proposed by Dijkstra and van Heuven (2002), assume that bottom-up processes govern the earliest stages of word recognition, increasing the activation of all form-related codes regardless of the language that is actually presented.

Perhaps the most straightforward way to investigate language selection during comprehension is to ask whether bilinguals exploit cues to selectively attend to the language in use. If language cues enable selection to occur in advance or to reduce cross-language activation that is in process, then bilinguals should be able to function as monolinguals. That is, the presence of cross-language ambiguity should have little consequence for processing. For example, a Chinese–English bilingual reading text in English, knows immediately that the text is not in Chinese and, in theory, should be able to shut down the activation of phonology or meaning that is shared across the two languages. Likewise, the context in which a bilingual is processing spoken or written information should provide higher level cues to better enable the selection of the intended language. The evidence, overall, provides surprisingly little support for the notion that bilinguals can utilize cues of this sort to select the intended language in advance. The situation is counterintuitive given the great number of potential language cues. But there are also some differences in the evidence for the comprehension of written text vs. spoken language and we consider below how the modality of language processing may affect language selection.

21.2.1 The search for cues: modulating non-selectivity in comprehension

A logical approach in the search for language selection in comprehension is to determine whether there are cues in the input or in the context of language use that can modulate markers of language co-activation. If a cue
is successful in reducing influence of the unintended language or enhancing the activation of the intended language, then the presence of that cue should eliminate cognate and homograph effects. Language cues can fall into at least two different categories, although these categories may not be mutually exclusive. The first category consists of factors that function as a cue because they exploit cross-language differences. The Chinese–English reader encountering text in English is an example because of the difference in written script between Chinese and English. One need not be bilingual to know that the English text is not Chinese but the question here is whether the bilingual can use that information to guide processing. The second category includes cues that capitalize on higher order contextual or linguistic information that may provide a pointer to the upcoming language. If bilinguals can exploit the presence of contextual cues, then processing text or listening to speech in one language alone should provide sufficient information to restrict processing to that language. As we will show, although there are some examples of language selectivity (e.g., Ju & Luce, 2004), more of the evidence demonstrates that bilinguals have difficulty in using this information to restrict processing (e.g., Van Assche et al., 2009).

21.2.2 Can bilinguals exploit language-specific features?
Examples of cross-language differences abound during language comprehension. One difference that should theoretically influence selection in word recognition includes the distinct lexical form of words in each language. While cognates and homographs share lexical form across the two languages, they also exhibit subtle differences in their degree of form overlap. The phonology of the bilingual’s two languages is never identical and the orthography can overlap to varying degrees. In the case of different-script bilinguals, there may be phonological overlap between cognates or homographs but no overlap in the orthography. In the case of bimodal bilingualism, where one language is written or spoken and the other signed (e.g., American Sign Language and English), the two languages are structurally distinct. Yet in each case, there is evidence that readers activate information, including the translation equivalent, in one language when reading the other (e.g., Morford et al., 2011; Thierry & Wu, 2007). Although the magnitude of cognate and homograph effects in languages that share orthographic and phonological codes appears to be modulated by the degree of overlap (e.g., Dijkstra, Grainger, & van Heuven, 1999; Schwartz, Diaz, & Kroll, 2007), in language pairings that do not share lexical form, distinct cross-language features do not function as a cue to enable language selective processing, at least when information is presented in written form.

The evidence on spoken word recognition is more mixed with respect to cue sensitivity than the evidence on visual word recognition. Some studies
have reported a pattern that is consistent with the language non-selectivity revealed in visual word recognition. Lagrou, Hartsuiker, and Duyck (2011) compared the performance of Dutch–English bilinguals on an auditory lexical decision task in which they had to decide whether spoken tokens were real words in one language alone. The tokens included Dutch–English homophones, words that sound similar but that have different meanings in the two languages. The cue that was introduced was the accentedness of the speaker’s voice, with either a native English or native Dutch speaker producing the tokens to be judged. Lagrou et al. found that bilinguals were indeed sensitive to the presence of accentedness overall, but the accentedness of the spoken words did not affect the magnitude of the homophone effect. Bilinguals were slower to judge homophones than control words regardless of whether the accentedness of the spoken voice provided a cue to the language spoken. The overall sensitivity to accentedness suggests that they were not deaf to the cue – they indeed processed it. But they did not use it selectively to interpret the input as coming from one language alone.

Other studies of spoken word recognition have used the visual world paradigm (Allopenna, Magnuson, & Tanenhaus, 1998) to examine cross-language effects. In the visual world paradigm, participants see a real or computer-generated grid of objects in front of them while they hear a spoken word. Their task is to point to or click on the object that is named by the spoken word. The critical manipulation is whether among the objects in the visual display, there is another object whose name is phonologically similar to the target name (e.g., candy–candle). Eye movements are tracked while the task is performed and the basic finding is that listeners fixate longer on objects whose names are phonological competitors with the target name. Marian and Spivey (2003b) reported that there is an effect of phonological competitors not only within languages, but also between languages when bilinguals perform the task in one of their two languages. Like other word recognition results in the visual domain, these cross-language effects are typically larger and more robust when performing the task in the L2, but the fact that they are observed when performing the task in one language alone reveals the fundamental non-selectivity of word recognition.

Subsequent studies using the visual world paradigm have shown that subtle phonetic differences can alter the selectivity of the system. For example, Ju and Luce (2004) found that language-specific differences in voice onset time (VOT) changed the degree to which participants fixated on cross-language homophone distractors. When Spanish words were pronounced with a Spanish VOT, English competitor items were not fixated on any more than compared to control words. In contrast, when Spanish words were pronounced with English VOTs, English distractors were then considered by the participants as evidenced by an increase in the number of fixations compared to controls. This finding suggests that Spanish
phonetic differences may have allowed the participants to selectively access Spanish without influence from English (see also Weber & Cutler, 2004). On the surface, the selectivity imposed by the presence of language-specific cues to the phonology in the visual world task would seem to be at odds with the lexical decision results of Lagrou et al. (2011) although the recognition tasks performed in each case are somewhat different. This leaves open the possibility that some types of cues may allow for selective access but it is not yet clear under what circumstances or over what time course.

In addition to orthographic and phonological ambiguity, language ambiguous words such as homographs or homophones frequently differ in their grammatical properties within each language. It appears that bilinguals may be able to capitalize on these differences in word class to exploit them as a language cue. Sunderman and Kroll (2006) found that lexical form interference could be eliminated in a translation recognition task (i.e., decide whether two words are translations of one another) when the two words differed in their grammatical class. When two words were not translation equivalents but the translation of one resembled the other (e.g., hombre (man)–hambre (hunger) in Spanish) there was interference only when the grammatical class of the two words matched. Likewise, Baten, Hofman, and Loeyes (2011) showed that word class modulated the degree of language co-activation in a lexical decision task. They found a facilitatory homograph effect in lexical decision only when the meaning of the homograph shared grammatical class with its translation. In each of these examples, it is not clear whether the locus of selection occurs early or late in processing. According to models such as the BIA+ (Dijkstra & van Heuven, 2002), the earliest stages of bilingual word recognition should be data driven and higher-order cues, such as grammatical class, should only influence selection processes at a relatively late point in processing. In future research it will be critical to use methods such as eye tracking and event-related potentials (ERPs) that permit a sensitive analysis of the early time course of processing to answer this question. However, like the VOT results in the visual world studies, the report of a pattern of processing that is influenced by grammatical class also suggests that bilinguals may be able to exploit cues to language status under some circumstances.

21.2.3 Can bilinguals use context to guide language-specific lexical access?

Words do not typically appear in isolation, they are embedded within multiple contexts: in sentences, in text, and in extended discourse. Much of the research on language selection in bilingual word recognition has focused on how different contexts of use (particularly list composition and sentential context) might influence non-selectivity. Cognate facilitation seems to be relatively robust in the face of different language contexts
Dijkstra et al. showed that cognate effects were present in an isolated lexical decision task regardless of the specific list composition, whether the lists contained words all in one language, or whether the list contained other-language distractor words. This was among the first studies to suggest that top-down contextual information was not utilized by bilinguals to control activation of the two languages.

With an isolated presentation, it is perhaps not surprising that both languages will be activated, especially for words with a high degree of cross-language overlap, such as cognates. In theory, isolated presentation heightens the ambiguity of language membership of a word. A Spanish–English bilingual who sees the cognate word “bus” presented alone would not know whether to activate Spanish or English, unless it were possible to utilize top-down information about the language membership of other words in the list. However, everyday language use includes many discourse and sentential features that go beyond list composition and that might function as language cues. A question then is whether a richer linguistic context, such as a sentence context, might provide bilinguals with information to predict the language of the upcoming words.

Quite surprisingly, even for words embedded in unilingual sentences, there is overwhelming evidence that bilinguals activate both language representations (e.g., Libben & Titone, 2009; Schwartz & Kroll, 2006; van Hell & de Groot, 2008). Duycket al. (2007) showed that when Dutch–English bilinguals read English sentences naturalistically cognate words were fixated on for less time compared to matched control words. This pattern of fixation differences originated during very early measures of reading time and persisted until later measures, suggesting that both languages were activated throughout the entire time course of processing, despite the presence of a coherent sentence context all in one language. When bilinguals recognize words within sentences, cognate and homograph effects are still present. In visual word recognition within sentence context, eye movements have been used to track lexical co-activation throughout the entire time course of recognition (e.g., Libben & Titone, 2009). Data indicate that both languages are activated from the initial stages of word recognition that track orthographic encoding and lexical access, and they remain active until later stages that include semantic integration. This suggests that the presence of a coherent sentence context by itself does not cause lexical access to become language selective. Indeed, a few studies have shown that these cross-language effects can be observed even when bilinguals are reading in the native language alone (e.g., Van Assche et al., 2009). The only exception to language non-selectivity has been reported in the presence of sentences that are highly semantically constrained, with highly constrained contexts reducing or eliminating language non-selectivity (e.g., Libben & Titone, 2009; Schwartz & Kroll, 2006; van Hell & de Groot, 2008; but see Van Assche et al., 2010).
Another potential cue is in the context of usage. Socio-contextual situations which require only one language might allow bilinguals to selectively access that language whereas situations in which two languages are required might show evidence of co-activation. This finding would be consistent with Grosjean’s (2001) language mode hypothesis, the idea that the activity of the bilingual’s two languages varies along a continuum from monolingual mode, with one language used primarily, to bilingual mode, where both languages are actively used. More recently, Green and Abutalebi (2013) have extended this hypothesis into the Adaptive Control Hypothesis stating that the context in which bilinguals acquire and speak their second language may fundamentally determine how language selection proceeds. Hence, different language contexts (e.g., unilingual context vs. code-switching context) are proposed to lead to distinct patterns of language selection and to place differential demands on the cognitive and neural processes that support each language. Surprisingly, research on word recognition in isolation and in sentence context suggests that context plays less of a role in initial language selection than might be expected. For example, Gullifer, Kroll, and Dussias (2013) showed that cognate effects were present and of the same magnitude for words recognized within sentences regardless of whether the language of the sentence context was consistent within the experiment or whether the language of the sentence context switched every two sentences. The magnitude of the cognate effect for cognate translations embedded within the sentence did not depend on whether the previous sentence was in the same or in a different language and there was virtually no switch cost to process target cognates in sentence contexts that had changed language from those that had not.

In comparison to cognates, homographs appear to be more sensitive to the language context. After sufficient experience with a unimodal context in an experiment, bilinguals show effects of “zooming in” to the intended language. Following sufficient exposure to the language of the sentence context, homograph effects are diminished or eliminated (e.g., Elston-Güttler, Gunter, & Kotz, 2005). Likewise, in research on isolated word recognition, the magnitude of homograph effects are more closely related with aspects of list composition. Homograph effects are greatest, for example, when other language distractor items are included in the list of words for a unilingual lexical decision task (Dijkstra et al., 1998). It may be the case that homographs simply offer a more sensitive measure of relative activation levels of each language and the control mechanisms that regulate those activation levels compared to cognates. In a dynamic system that is under the control of executive function mechanisms and must take into account the context of language usage the contrast in sensitivity to context makes sense. Parallel activation of cognates is almost always beneficial to processing providing a facilitatory effect, where the parallel activation of homographs imposes a conflict that is only occasionally
beneficial to processing in the face of conflict in meaning across languages. Hence, for homographs, bilinguals who are particularly skilled in control of both languages may be able to regulate the inhibition of competitors that are detrimental to fluent processing. By comparison, the resonance created by the convergence across lexical codes for cognates may be driven more directly by bottom-up processes that are immune to the influence of control mechanisms (for additional evidence on access in sentence context, see Titone et al., 2011).

The persistence of language non-selectivity, even in the presence of rich context, is counterintuitive and puzzling. Indeed, the evidence reviewed earlier, suggesting that there are no switch costs in lexical access when the entire sentence context switches back and forth, is surprising. But what happens when words switch from one language to the other? Language switching has been used as a means to understand the nature of the selection mechanism. Lexical switch costs have been documented in a number of word recognition studies (e.g., Thomas & Allport, 2000). Typically a switch cost is observed, with longer response times on trials following a language switch than following a same-language trial. That result is notable because the persistent activity of the two languages that we have documented might have led to a prediction of no switch costs if the bilingual is always in a functionally mixed language context. Critically, Thomas and Allport found effects of language-specific orthography but no modulation of those effects in the switch and no-switch conditions, suggesting that language selection affects late stages of processing in word recognition. As we will see in reviewing the evidence on bilingual production, the pattern of switch costs in comprehension stands in contrast to the one reported for production. While the explicit mechanism of language selection during comprehension is yet to be unequivocally identified, there is emerging research elucidating the link between a bilingual’s capacity for inhibitory control and language selection and co-activation during comprehension. We turn to that evidence next.

21.2.4 How do bilinguals control selection processes in comprehension?

The fact that cross-language activation is observed even when bilinguals are required to only use one of their languages raises the question about how they select the language they need according to the context. Much of the recent research on bilingual language processing has focused on understanding the control mechanisms that allow them to overcome the negative influence of activating their two languages, and importantly, how the consequences of this parallel activation impact domain-general cognitive processes.

Two important remarks are relevant for language selection: first, given that language co-activation occurs, the presence of between-language
competition introduces the need for a control mechanism that regulates the activation of the non-target language in order to correctly select the intended one; second, the non-selective activation of the two languages may give rise to differences in lexical access and the degree of between-language competition that depend on several factors that influence bilingual processing (Green & Abutalebi, 2013). For example, lexical selection may vary depending on the locus of cross-language activation, and so cross-language interference and the control effects may occur at different levels of language processing (Kroll, Bobb, & Wodniecka, 2006). Considering this, it is important to note that the consequences of the cross-language activation might be not the same for comprehension and production. Production is a conceptually driven process, in which the concept of a given word is activated before the word form and phonological properties of possible target words are available (e.g., Levelt, 1989). Comprehension, as we have documented in the studies we have reviewed above, is a bottom-up process, in which the orthography and phonology drive later conceptual access (e.g., Dijkstra & van Heuven, 2002). The implication of these fundamental differences between production and comprehension is that the locus and the control mechanisms for lexical selection may differ across both domains.

While there is no agreement about the nature of the control mechanisms that enable language selection, a recent body of evidence suggests the presence of inhibitory processes (for a review, see Kroll et al., 2008). Although most of the empirical evidence supporting inhibitory processes in bilingual processing comes from the language production domain (Costa & Santesteban, 2004; Meuter & Allport, 1999) recent studies also demonstrate the involvement of inhibitory processes in bilingual language comprehension (Macizo, Martín, & Bajo, 2010).

As reviewed in the earlier section on language non-selectivity, most of the studies on cross-language interactions have used the general strategy of examining words that share lexical, orthographic, or phonological properties in two languages (e.g., false friends or homographs, cognates, and homophones). So, for example, in visual and spoken word recognition, competition may arise among lexical neighbors with either similar orthography or phonology. Following this strategy, Martín, Macizo, and Bajo (2010) investigated the role of the control processes involved in language selection using a paradigm that allows observing the cross-language activation and how it is resolved. They asked Spanish–English bilinguals to perform semantic relatedness judgments to English word pairs. On the critical trials, interlingual homographs were presented among the word pairs (e.g., the word “pie” means foot in Spanish and cake in English). On the following trial, the word pair included the English translation of the Spanish meaning of the homograph. Bilinguals were slower to judge word pairs that contained a homograph (e.g., pie-toe). Critically, after responding to homographs, bilinguals slowed their responses when its Spanish
meaning became relevant in the subsequent trial as compared to a control word pair (e.g., foot–hand). The interference effect found in the first trial was taken as an index of cross-language activation. Given the interference effect stemming from the parallel activation of the two homograph meanings, the bilinguals appeared to resolve the interference by suppressing the non-target and competing homograph meaning in order to select the appropriate one.

Martin et al. (2010) focused on the interval between the time in which interference was produced and the presentation of the irrelevant homograph meaning as a means to map out the time course of resolving inhibition. They found that the inhibitory effects were relatively short lived since they were observed in a time interval between 500 ms and 750 ms. After an interval of 750 ms, they seem to have recovered from inhibition. These findings indicate that inhibitory control processes are involved in the resolution of cross-language competition in comprehension, and this inhibition has transient effects. A critical point is that overcoming inhibition may take time and it can impose a cost in bilinguals' performance. The effects we have described were observed when bilinguals processed words, out of context, and under artificial experimental conditions. Bilinguals rarely use their two languages in this manner. In future work it will be important to ask whether these time-constrained processes are manifest when bilinguals comprehend language in higher-level context and, if so, whether context modulates the presence and time course of language selection.

Converging support for the idea that there is a brief period of inhibitory control that can be caught on the fly in comprehension has been reported by Blumenfeld and Marian (2011). They developed a variant of the visual world paradigm that has been used extensively as a means to investigate spoken word recognition. On a first trial, native English-speaking bilinguals and monolinguals were presented with a display in which they had to identify a word in English, the L1 for both groups. Their eyes were tracked while they performed the identification task. One of the pictures displayed on the first trial had a name that was a phonological competitor to the target English word that was spoken. Blumenfeld and Marian reported that both bilinguals and monolinguals experienced within-language competition on the first trial, revealed in the eye movement record, with somewhat longer fixations on the phonological competitor than on the control pictures. Critically, on a subsequent trial, the same grid was presented but without pictures of objects. Instead, in one of the four corners of the grid, an asterisk appeared and the participant was asked to click on the asterisk. They found an inhibitory pattern for the monolinguals on the second trial when the asterisk appeared in the position on the grid that had previously held the phonological competitor. Blumenfeld and Marian argued that the bilinguals had apparently resolved the inhibition by the time of the second trial, whereas the monolinguals had not.
They suggest that bilinguals are more efficient at resolving cross-language competition online, a result that is congenial with the findings of brain imaging studies that show that bilinguals produce less activation than monolinguals in brain areas such as the anterior cingulate cortex (ACC) that are implicated in the resolution of competition in non-linguistic tasks. Without a fine-grained analysis of the time course of processing, it is difficult to compare the Blumenfeld and Marian results directly to those of Martín et al. (2010), and of course the tasks are quite different, spoken word recognition in one case and a semantic relatedness judgment in the other. However, in both studies there is evidence that suggests that bilinguals cannot turn off the initial bottom-up processes that produce competition in comprehension. Rather, they learn to modulate its consequences in the moments afterwards. Each of these studies also suggests that the processing of inhibiting competitors is one that extends over a relatively brief time-span in comprehension. As we will see, this will stand in contrast to the findings in production, where inhibitory processes may include longer-lasting consequences for performance.

21.3 Bilingual production

Like the evidence on comprehension, studies of bilingual speech planning show that information about both languages is active, at least momentarily, when bilinguals plan to speak even a single word in one language alone (e.g., Costa, 2005; Hanulová, Davidson, & Indefrey, 2011; Kroll et al., 2006). The parallel activation of two languages occurs for speakers who are highly proficient in both languages as well as for those who are still learning the L2. A focus in the most recent research in bilingual language processing is to understand what mechanisms allow bilinguals to negotiate language activation and, once proficient, to make few language errors (e.g., Gollan, Sandoval, & Salmon, 2011). Although bilinguals normally make few errors of language, the importance of a control mechanism has also been recognized in cases of aphasia where bilingual speakers who suffered neurological damage cannot properly control language selection, leading to pathological language mixing (Abutalebi & Green, 2008).

21.3.1 Two views of bilingual speech planning

A number of different mechanisms have been hypothesized to allow bilinguals to constrain selection to the intended language during speech planning (for a recent detailed review, see Kroll & Gollan, 2014). Findings also generally indicate that parallel activation can have multiple consequences, making the locus of language selection in production variable rather than fixed (e.g., Kroll et al., 2006), and dependent on a range of factors including the context of the language to be spoken (e.g., language
immersion and the language profile of interlocutors), proficiency in the L2, and the demands of speaking associated with the particular production task to be performed. One alternative proposes that selection is language specific (e.g., Costa, Miozzo, & Caramazza, 1999; Finkbeiner et al., 2006; Finkbeiner, Gollan, & Caramazza, 2006). On this view, there may be activation of words within the language not in use but the activation of those words does not make them candidates for selection. This first hypothesis would require that bilinguals have a mechanism in place that allows them to use cues to enable them to select the language as efficiently as possible, a solution that we have called the “mental firewall” (e.g., Kroll et al., 2012).

Like the research on bilingual comprehension, external cues may be related to linguistic features of the two languages (e.g., phonological or lexical cues) but also to features of the context in which the two languages were acquired and are used. However, studies of language production that have examined the role of language cues, such as cross-language script differences, have failed to provide evidence that bilinguals can easily exploit the available cues (Hoshino & Kroll, 2008). Similarly, typological differences that should provide a clear means to categorize the two languages do not appear to function effectively as cues. For example, production studies in hearing bimodal bilinguals who speak one language and sign the other, have also shown that there are cross-language influences even when production engages different articulatory systems (e.g., Emmorey et al., 2008). It is possible, however, that script and typological differences are subtle in the sense that activation of information pertaining to script or typology does not occur in response to a trigger in the immediate environment. Cues may only become effective as a consequence of the bilingual’s lifetime experience, with collective past experiences likely to be stored in memory (Jared, Poh, & Paivio, 2013), but not accessible in the moments leading up to speech unless there is an additional contextual trigger (e.g., Zhang et al., 2013). In this sense, the evidence on production would seem to mirror the results we have reviewed for comprehension. In each case, there appears to be activation of alternatives related to both the target language and the language not in use. For production, however, this observation is quite counterintuitive, because unlike comprehension, production is initiated by a top-down process that first engages ideas and then maps them to their respective linguistic forms. Logically, it would seem possible to identify the intended language at an early stage of speech planning.

If the logic of the language selective view is straightforward, providing a clear explanation of how selectivity is accomplished is not. By a language-selective account, there would have to be a means for sending greater activation to the intended language and the studies to date seem to suggest quite clearly that the intention to plan speech in one language alone is insufficient to restrict activation to that language (but see La Heij, 2005, for a defense of the selective position).
The alternative view is that all activated candidates compete for selection. Superficially, this view would seem to align with the models of word recognition like BIA+ that account for non-selectivity in comprehension by assuming that bottom-up activation spreads to form-related words in both languages. But it is important to remember that the top-down planning of speech means that the codes that are first activated are related in meaning rather than form with the ideas to be communicated. Those meanings may subsequently activate word forms, including translation equivalents and phonological relatives (e.g., Colome & Miozzo, 2010; Hoshino & Kroll, 2008; Hoshino & Thierry, 2011), but the sequence of processing differs in comprehension and production. Although the initiation of speech planning begins at a conceptual level, the evidence suggests that cross-language activation reaches all the way to the phonology and even beyond, to the execution of speech that is manifest in the acoustic representation of the spoken utterance, to the point where the unintended word may literally be on the tip of the bilingual’s tongue (e.g., Gollan & Goldrick, 2012).

### 21.3.2 Evidence on language switching and language mixing

Studies of language production often adopt a logic similar to the one applied to word recognition so that bilingual speakers are asked to produce language ambiguous words, such as cognates. In production, unlike comprehension, the cognate word is not actually present, so showing that there is facilitation in naming a picture whose translation is a cognate in the language not to be spoken provides evidence for cascading activation all the way to the phonology of the alternative (e.g., Costa et al., 2000; Hoshino & Kroll, 2008).

In comprehension, the effects of cross-language ambiguity are perhaps the most compelling in revealing the architecture of the system because language ambiguous forms affect the earliest stages of recognition. In production, the evidence for language non-selectivity has relied more on the consequences of language switching and/or mixing. The logic is simple. If bilinguals can plan speech in one language alone, as if they were monolingual speakers, then a cost should be observed when there is uncertainty about the language to be spoken or when the languages switch from one utterance to the next. For production, this logic affects the earliest stages of planning, potentially revealing the most basic features of the architecture that underlie speech, unlike cognate facilitation effects that necessarily reflect later processing once speech planning has already been initiated.

---

2 A third alternative is the frequency lag or weaker links hypothesis that has been proposed by Gollan and colleagues (e.g., Gollan et al., 2008) to account for the relatively lower frequency of use of each language in bilinguals relative to monolinguals. See Kroll and Gollan (2014) for a discussion of how this alternative contrasts with the competition for selection hypothesis.
21.3.3 Costs to language switching and mixing

When bilinguals are forced to switch from one language to the other in a naming task (e.g., name pictures or digits) there is a cost to processing, with longer naming latencies following switch than no-switch trials (e.g., Meuter & Allport, 1999). In language switching tasks, a picture or number is presented with a cue signaling in which language the stimulus is to be named. The finding that there is a cost to naming following a language switch might be interpreted as support for a language selective model of bilingual production. However, Meuter and Allport and many subsequent studies have reported that there is a curious asymmetry in switch costs, with larger costs when switching from the L2 into the L1 than the reverse. The greater switch cost into the L1 was initially taken as support for the claims of the Inhibitory Control (IC) model (Green, 1998) which suggested that production of the L2 requires inhibition of the L1 and that subsequent production of L1 then reflects the spillover of that inhibitory process. It is beyond the scope of the present chapter to provide a comprehensive review of the switch cost asymmetries and there are other many discussions of how these asymmetries might be interpreted (e.g., Bobb & Wodniecka, 2013; Gollan & Ferreira, 2009; Schwieter & Sunderman, 2008). For present purposes, the critical observation is that there are switch costs and that sometimes the more dominant L1 reveals those costs more clearly than the less dominant L2.

A similar differential pattern has been reported for language mixing (e.g., Kroll et al., 2000). When the language of production is uncertain, the L1 suffers a cost and that cost can be observed to the point where even highly L1-dominant bilinguals are slower to speak the L1 than the L2 and where the consequences of mixing can be observed not only in behavior but also in the earliest time course of planning revealed using ERPs and in brain imaging patterns using fMRI (e.g., Christoffels, Firk, & Schiller, 2007; Guo et al., 2011; Kroll et al. (2000) found that Dutch–English speakers were slower to produce the Dutch names of pictures when faced with the requirement to produce picture names in English unpredictably relative to blocked trials in which they produced words in one language only. Critically, the time for the same speakers to produce English L2 names of pictures was unaffected by whether picture naming was blocked or mixed, suggesting that the L1 is active during speech planning in the L2, regardless of whether there is a requirement to make it active.

A language switching study that has generated a great deal of discussion was reported by Costa and Santesteban (2004). They compared the language switching performance of balanced and highly proficiency Spanish–Catalan bilinguals with less proficient speakers for whom there was clear dominance in the L1. For the L1-dominant speakers, they replicated the switch cost asymmetry reported by Meuter and Allport (1999) but for the balanced and high proficiency speakers, the switch costs were symmetric.
Selection and control in comprehension and production

for the two languages. The different pattern for the two speaker groups was initially interpreted as meaning that inhibitory control is required for those who are less proficient but once high proficiency is achieved, there is no longer a need for inhibition. Essentially the idea was that once the two languages are equal sparring partners, they can be engaged without differential regulation. However, another result in the Costa and Santesteban study betrays that interpretation. Although there was indeed a difference in the symmetry of the switch costs for the different types of bilingual speakers, there was also evidence that even for the high proficiency and balanced bilinguals, under the conditions of language mixing that are required by the switching paradigm, the L1 was produced more slowly than the L2. The fact that these two sources of evidence on switching and mixing produce different results can be understood if we assume that in production there may be multiple components of inhibitory control (for a discussion of these issues in the domain of translation and interpretation, see de Groot & Christoffels, 2006). The recent research on bilingual production has focused on this issue as a means to understand how language selection might be achieved.

21.3.4 Components of inhibitory control in production

The forced switching that is imposed by the language switching paradigm requires that bilinguals speak words in each of the two languages in a sequence that is highly unnatural. Although one might argue that all decontextualized lexical performance is unnatural to some degree, bilinguals often switch between two languages after speaking one language for a period of time. Another approach to investigating selection in production was taken by Misra et al. (2012). The idea was to enable bilinguals to speak one language alone, the way that they might more naturally, and then to impose a switch of language across blocks. If the switch costs that have been documented in forced trial-to-trial switching reflect a process that is imposed by the artificial nature of the task, then within a few trials, bilinguals should recover from the switch. Misra et al. asked relatively proficient Chinese–English bilinguals to name pictures in blocks of trials in either L1 (Chinese) or L2 (English). The manipulation was only whether naming was performed first in the L1 and then in the L2 or in the reverse order. The pictures to be named in each language were identical, with the prediction of priming on the second presentation of a picture relative to the first. Misra et al. recorded ERPs while the pictures were named and also behavioral measures of response time and accuracy. They found that when pictures were named in the L2 following the L1, the hypothesized priming was observed. ERPs were less negative on the second naming of a picture than the first, suggesting that there was facilitation in processing. In contrast, an inhibitory pattern was observed when the L1 followed the L2. Not only was there not a pattern consistent with priming, but there was
a larger N2 component and greater negativity for the L1 following the L2, consistent with the presence of inhibition for the L1. Beyond the different patterns as a function of language naming order, the most surprising result was that the inhibitory pattern in L1 following L2 was maintained over the course of two blocks of L1 naming. These bilinguals had ample opportunity to recover from any momentary inhibition imposed by speaking the L2 but many trials later were still showing the consequences (and see Phillip, Gade, & Koch, 2007, and Philipp & Koch, 2009, for evidence from the N-2 repetition paradigm that supports a similar inhibitory account). The use of the same pictures across languages in the Misra et al. study did not permit an assessment of the scope of inhibition but the pattern suggests that there is inhibition that is extended in time. As noted earlier, there may be multiple components of inhibitory control. Local inhibition may operate over specific words or conceptual categories and may extend for brief periods of time. Global inhibition may engage an entire language and last for a relatively long time.

Guo et al. (2011) used fMRI to track the patterns of brain activation when bilinguals name in extended blocks as they did in the Misra et al. (2012) study. The design in the Guo et al. study was identical to the blocked picture-naming procedure in Misra et al. except that following the blocked picture-naming trials, Chinese–English bilinguals named pictures in a mixed language block. Critically, different patterns of brain activation were found when comparing the blocked and mixed naming trials (hypothesized to reflect local inhibition) and the spillover effect of naming in different block orders (hypothesized to reflect global inhibition). The dorsal anterior cingulate cortex (ACC) and the supplementary motor area (SMA) appeared to play important roles in local inhibition, while the dorsal left frontal gyrus and parietal cortex appeared to be important for global inhibition.

In recent studies, each of these observations has been replicated in behavioral studies that used different tasks and different language pairings. For example, Van Assche, Duyck, and Gollan (2013) compared performance on a verbal letter fluency task as a function of the order of the language blocks (L1 or L2 first), the type of speaker (Dutch–English or Chinese–English), and whether the letter cues were the same or different across languages. They replicated the block order effect reported by Misra et al. (2012) in that letter fluency was reduced for the dominant language when it followed the less dominant language. The groups differed, however, in that only the Chinese–English bilinguals, but not the Dutch–English bilinguals, showed these effects globally, regardless of whether the letter cues were repeated or not. The pattern of results suggests that all bilinguals, regardless of their proficiency, show evidence of inhibitory processing, but that proficiency may determine the scope of inhibition.

Martin, Bajo, and Kroll (2013) reported a study that compared the behavioral performance of relatively proficient Chinese–English bilinguals
when they named pictures in blocked or mixed language trials. The innovation in this study was to add a concurrent updating task that required speakers to listen to a continuous series of tones and to press a key whenever three equal tones appear consecutively. The logic here was to determine whether the addition of a dual task would selectively disrupt inhibitory processing. The results showed that the updating task eliminated the block order effect. When bilinguals named pictures in L1 following L2 while they were performing the dual task, there was no apparent inhibition of the L1. In contrast, a robust effect of language mixing was observed regardless of whether bilinguals were performing the updating task or not. Although bilinguals were slower to name pictures overall when performing the concurrent updating task than not, the result of interest was that the disruptive effects were differential for the two hypothesized components of inhibitory control.

The findings on bilingual production converge on the conclusion that bilinguals, regardless of their proficiency, inhibit the L1 to enable speech planning in the L2. The clue first present in the Costa and Santesteban (2004) language switching data, that there might be more than one component of inhibitory control, has been supported by the recent studies that were designed to examine this issue explicitly. It remains to be determined how not only language proficiency and dominance, but also the context of language use, may modulate each of these components. Green and Abutalebi (2013) argue that the neural networks that support bilingual language processes are necessarily tuned differently in response to the requirement to engage these processes differentially. So a bilingual who is a habitual code switcher and living in an environment in which code switching is prevalent, may engage inhibitory mechanisms differently than a bilingual who uses each of his or her two languages in separate environments. Likewise, individuals who are immersed in an L2 environment may adjust the need for endogenous control. Linck, Kroll, and Sunderman (2009) showed that university students studying abroad for a semester produced reduced output in their L1 in a category fluency task relative to classroom learners but their L1 performance rebounded upon their return home (and see Baus, Costa, & Carreiras, 2013, and Levy et al., 2007, for a related laboratory version of language immersion). Critically, what these new data on language production show is that inhibitory control processes are engaged by the most proficient bilinguals as well as by L2 learners. It remains to be seen which of these processes are ephemeral, producing short-term effects that are modulated by context and language usage, and which of them depend on characteristics of the bilinguals themselves and the linguistic structure associated with the bilingual language pairings (see Lev-Ari & Peperkamp, 2013, for a recent discussion of this issue for how individual differences in inhibitory control may modulate the effects of the L2 on L1 phonetics).
21.4 Language selection in comprehension vs. production

The research we have reviewed in this chapter demonstrates that language selection is not a simple nor unitary mechanism. Bilinguals develop the means to negotiate persistent cross-language activation and to resolve the resulting competition in ways that reflect the demands that are selective in comprehension and production. In comprehension, there appear to be short-lived inhibitory processes whereas in production there may be control processes that require different types of inhibition and that sometimes extend over relatively long periods of time. There is a great deal that remains to be investigated, including the implications of the selection mechanisms in comprehension and production for domain-general cognitive performance and the way that these language processes, in and of themselves, may change and adapt in different circumstances.

Green and Abutalebi (2013) argue there may be particular language skills and contexts that determine how the neural networks that support language change with experience. A topic that we haven’t mentioned in our review but that draws on both comprehension and production in unique ways is translation and simultaneous interpretation. In a sense, skilled translation is a special form of extreme bilingualism. For interpreters, in particular, there are time pressures that render the task of using both languages quite differently than ordinary bilingualism. Although the basic research program on these issues is relatively new, the picture that emerges suggests that these skills sometimes modulate basic mechanisms of language processing and sometimes do not. For example, in a study examining lexical processing in sentence context, Ibáñez, Macizo, and Bajo (2010) reported that translators were less likely to engage inhibitory mechanisms. That finding makes sense, from the perspective that the goal for translators is to activate the other language continually, but demonstrating that it affects their performance even when they are not actively translating suggests that the nature of their language experience has consequences that extend beyond specific tasks. Likewise, Christoffels, de Groot, and Kroll (2006) showed that although interpreters have exceptional memory skills, a finding that in and of itself is difficult to interpret with respect to cause and effect, their performance in basic lexical production tasks like those we have reviewed in this chapter, is similar to other high proficiency bilinguals. In this instance, their cognitive abilities did not appear to modulate language processing.3

3 In the literature on translation and simultaneous interpretation there is often discussion about the contribution of possible self-selection factors. Only particular people seek training in these skills and it is not clear whether the observed cognitive advantages reflect individual differences that led them to seek training or the consequences of the training itself. Likewise, in a context such as the Netherlands, where virtually all university-educated individuals are multilingual, it is not clear whether the absence of differences on basic language processing tasks for interpreters and ordinary bilinguals means that there is no consequence of having these skills for language processing or whether the bilinguals to
The findings of the studies reported above suggest that the performance of professional translators and interpreters might not differ from other bilinguals while performing different production tasks but, at the same time, they show differences in the use of cognitive control in bilingual processing. In line with this source of evidence, a recent study shows that translators show a similar pattern of performance in comprehension tasks as that reported for language production. Martín et al. (under review) tested a group of professional translators using the same procedure as Macizo et al. (2010). The translators performed a semantic relatedness task in their L2 (i.e., English) on word pairs including interlingual homographs or control matched words (see the description of the paradigm in the section about selection processes in comprehension of this chapter). They found that the translators were similar to the control bilingual group, showing longer latencies in response to the word pairs containing homographs. Given that they showed the interference effect in this kind of trial, it would be expected that after responding to homographs the translators also performed similar to the control bilinguals showing an inhibitory effect in the subsequent trial as a sign of the suppression of the non-target and competing homograph meaning. However, unlike the bilingual control group, the translators did not show the inhibitory effect as a result of solving cross-language competition. This results show nicely how a specific bilingual context can shape the experience in the usage of control processes involved in language processing. In the practice of their profession, a translator rarely speaks. The input for a translator comes mainly from the comprehension domain and the output is elaborated in written production. Although they do not have the temporal pressure as interpreters do, they are required to change from a language to another constantly. It is not strange then, that when they are required to work only in one language, they perform as other bilinguals do. However, the particular experience they have changing between languages shows that the interplay between comprehension and production shapes the way in which they use cognitive control mechanisms and it is reflected in a differential performance solving cross-language competition (Green & Abutalebi, 2013). A question for future research will be to determine the scope of the consequences of particular types of bilingual language experience in both comprehension and production for altering the manner in which cross-language competition is negotiated.

One implication of the research that we have discussed is that bilinguals and monolinguals might be understood to be processing language under very different task demands and those differences might be expected to have consequences not only for language, but also for whom the interpreters were compared also use their languages in a unique environment that masks the contribution of other factors.
cognition and brain function. Recent imaging studies (e.g., Parker Jones et al., 2012) have reported differences in patterns of brain activation for bilinguals and monolinguals even when they perform the same tasks in their L1 only. Because the neural support for language processing is shared across the bilingual’s two languages (e.g., Abutalebi et al., 2005) and because the networks of control that are associated with them change with language experience (e.g., Abutalebi et al., 2012; Gold et al., 2013), it may not be surprising that differences between monolinguals and bilinguals emerge even when they perform the same tasks only in the native language.

The characterization we have provided of selection and control in bilingual comprehension and production might allow a reader to come to the mistaken conclusion that these processes are independent of one another. To the contrary, we assume that they are tightly linked (e.g., MacDonald, 2013). One context in which it is possible to begin to identify the way in which performance in production provides a cue to comprehension comes from studies of code switching behavior. Studies have begun to ask whether the few markers of disfluency that occur in the context of code switches are informative about the locus of planning and control in code switching mode. Hlavac (2011) found that Croatian–English speakers tended to produce few hesitation and monitoring phenomena (e.g., hesitations, pauses, verbal fillers) overall in code switched speech (around 2 percent of the total utterances contained these phenomena). He also found that a significantly greater proportion of verbal fillers preceded than followed a code switch. More importantly, when hesitation and monitoring phenomena occurred, they were found more often in the context of code switched words that do not contain the phonology or morphological patterns of the base language. The Hlavac results suggest that presence of disfluencies may not be the result of retrieval difficulties during code switches, but instead may serve to allow the speaker to signal an upcoming switch to the listener.

The pattern of code switching behavior reported by Hlavac can be understood within the framework of the Adaptive Control Hypothesis (Green & Abutalebi, 2013). The description of the highly proficient group Croatian–English speakers sampled for his study suggests that this group may have adapted to a language context in which dense code switching is the norm. The Adaptive Control Hypothesis predicts that these bilinguals have developed a strategy that allows the speaker to opportunistically exploit co-activation of two languages, rather than having them in competition. It is our interpretation that the use of integrated word forms in code switched speech is one sign that cooperative co-activation of L1 and L2 exists. Code switching may therefore provide a model, and a uniquely bilingual model, for evaluating the way in which comprehension is tuned to production.
21.5 Conclusions

In this chapter we reviewed the basic evidence for language non-selectivity at the lexical level in bilingual comprehension and production. At a general level, the findings in the two domains are similar. There is parallel activation of the bilingual’s two languages in both comprehension and in planning for production and it is difficult, if not impossible, for bilinguals to easily exploit available cues to the language in use to override language non-selectivity. At a more detailed level, comprehension and production differ in some important respects, including the nature of the information that is active and competing across languages, the scope of inhibitory processes, and the time course over which inhibition occurs. There is a call now to relate language processing more closely to its cognitive and neural consequences and to identify the causal basis of these experiential changes (e.g., Baum & Titone, 2014; Green & Abutalebi, 2013; Kroll & Bialystok, 2013). The research findings we have discussed provide an initial basis to pursue this question, paying closer attention to what it is that bilinguals are doing with language when they speak and understand one another. Like other proposals, our analysis of differences in comprehension and production suggests that the situation is more complex than we understood it to be. But it also shows that using bilingualism as a tool demonstrates the way that juggling two languages in one mind and brain may reveal more about foundational principles than we could ever know by studying monolingual speakers alone.