Title: The neuroanatomy of bilingualism: how to turn a hazy view into the full picture.

Running title: The bilingual brain: a critical review.

Authors: Lorna García-Pentón[a], Yuriem Fernández García[a], Brendan Costello[a], Jon Andoni Duñabeitia[a] & Manuel Carreiras[a,b].

Institutions:

[a] Basque Center on Cognition, Brain and Language (BCBL), Mikeletegui 69, 2º, 20009 Donostia-San Sebastian, Gipuzkoa, Spain.

[b] IKERBASQUE. Basque Foundation for Science, Alameda Urquijo, 36-5, 48011 Bilbao, Bizkaia, Spain.

Corresponding author: Lorna García Pentón

e-mail: l.garcia@bcbl.eu

Basque Center on Cognition Brain and Language (BCBL)

Paseo Mikeletegi 69, piso 2, 20009 Donostia, Gipuzkoa,

Spain.

Tel: +34 943 309 300

Fax: +34 943 309 052

Title: The neuroanatomy of bilingualism: how to turn a hazy view into the full picture.

Abstract:

The neuroanatomical bases of bilingualism have recently received intensive attention. However, it is still a matter of debate how the brain structure changes due to bilingual experience since current findings are highly variable. The aim of this review is to examine these structural studies from a methodological perspective and to discuss two major methodological problems that could give rise to this variability. The first problem is sample selection, an issue directly related to the heterogeneous nature of bilingualism. The second problem is the inconsistency in the methods used for the analysis of brain imaging data. This review reveals that although structural changes related to bilingualism have been reported in regions comprising language/cognitive control and language processing, these results are not yet sufficiently numerous or consistent to allow important generalizations to be reached. Consequently, current evidence offers ambiguous support for neural models of bilingualism. This shortcoming in the field is exacerbated by critical methodological differences between studies that only further complicate the matter. We conclude by identifying issues that should be taken into consideration so that studies are more comparable and results are easier to aggregate and interpret. We also point out future directions that would allow for progress in the field.

Keywords: bilingualism, neuroplasticity, VBM, DTI, second language learning.

1. Introduction

Critical brain areas related to language have been extensively studied and described (Price, 2010). However, the question of whether we need different or extra brain language regions/subnetworks to support more than one language remains controversial. The practical and theoretical repercussions of this issue make studies that investigate the biological basis of a second language (L2) particularly desirable (Kennedy & Norman, 2005). Two conditions make this line of research possible. Firstly, contemporary neuroimaging techniques and methods provide the tools to investigate second language processing. Secondly, the widespread existence of bilingual (and multilingual) populations throughout the world provides ample opportunity for looking into those brain changes related to the acquisition of additional languages.

In this day and age, bilingualism is a common reality for millions of people in the world: in the context of globalization, members of all societies are exposed to languages other than their own, with estimates that more than half of the world's population uses two or more languages (Grosjean, 2010) and two thirds of the world's children grow up in a bilingual environment (Crystal, 1997).

However, the way in which a second (or subsequent) language is acquired varies significantly across and within societies, ranging from individuals learning two languages with extensive contextual presence of both languages from birth (the case of simultaneous bilinguals in settings such as the Basque Country, Wales or Catalonia, among many others), to late learners of a second language (L2) with restricted or low contextual presence in the environment (e.g. learning a second language through classroom instruction without L2

natural immersion). These varying conditions for the acquisition of more than one language clearly impose sociolinguistic differences among bilingual individuals.

Recent years have witnessed an exponential increase in the number of research reports making claims for a generalized 'advantage' of bilinguals over monolinguals in tasks mainly tapping into attentional resources, memory skills and executive control mechanisms. In the current study, we will solely focus on the potential differences between bilinguals and monolinguals argued to exist in the brain circuitry, the neuroanatomical differences between bilinguals and monolinguals being our exclusive centre of interest. However, in order to give the reader an approximate idea of the state of the art in this matter, we will first refer to a specific domain in which the stability and replicability of the differences between these two groups have yielded lively debate recently: the bilingual advantage in executive functions (see Paap, Johnson, & Sawi, in press, for a comprehensive review). We note, however, that in this article we leave aside other potential differences that may emerge as a function of the knowledge and/or use of more than a single language, given that it is not the primary focus of this review. There is much evidence demonstrating that bilinguals outperform monolinguals in executive control tasks (see Abutalebi & Green, 2007; Bialystok, Craik, & Luk, 2012 for a review; see also Bialystok & Barac, 2012; Costa, Hernández, Costa-Faidella, & Sebastián-Gallés, 2009; Costa, Hernández, & Sebastián-Gallés, 2008). At a broad level, the underlying hypothesis for the so-called ‘bilingual advantage’ in executive functions is that bilinguals are used to constantly dealing with different languages and to preventing mutual interference between languages by selecting the target language while inhibiting the non-target language(s). This practice provides bilinguals with a somewhat enhanced mental flexibility, which results in augmented or

improved skills related to the management of conflicting information as compared to monolinguals (see Kroll & Bialystok, 2013, for a review). In other words, speaking several languages can lead to benefits that go beyond the realm of language, impacting on global cognitive functioning (Bialystok et al., 2012) and, more specifically, on the mechanisms responsible for selecting one language while managing interference from the other(s). These mechanisms are assumed to be "most likely found in the executive control system that is largely based on a network of processes in the frontal cortex" (see Kroll & Bialystok, 2013, p. 498), but there are other regions involved (see below for a detailed description of the brain network for language control). Findings suggesting that language control and cognitive control recruit similar neural mechanisms have been taken to support this hypothesis (see Abutalebi & Green, 2007; Luk, Green, Abutalebi & Grady 2012 for review). However, demonstrating that both language and cognitive control mechanisms overlap in a distributed fronto-parietal network (De Baene, Duyck, Brass & Carreiras, 2015) does not necessarily imply a bilingual advantage. Furthermore, finding differences between bilinguals and monolinguals in the recruitment of brain structures for tasks that require different sources of control (e.g., language control and non-linguistic interference control) does not directly speak to the existence of an advantage (see Duñabeitia & Carreiras, in press). Supporting evidence for an advantage should involve showing that these differences are accompanied by unambiguous behavioural data substantiating a cognitive gain or, for instance, a demonstration of how these differences mitigate normal or pathological cognitive decline in the elderly (see Clare, Whitaker, Craik, Bialystok, Martyr, Martin-Forbes et al., 2014, for review; see also Duñabeitia, Fernández, & Carreiras, submitted). Thus, showing more than neuroanatomical differences between bilinguals and monolinguals is needed to underpin any possible bilingual advantage.

The matter of the bilingual advantage[1] in the behavioural domain is not without its complications. Some of the recent behavioural studies investigating bilingual advantage tested vast numbers of simultaneous and early bilinguals and reported no difference between bilinguals and monolingual peers in tasks such as verbal Stroop, number size congruency (a non-verbal version of the Stroop) (Duñabeitia et al., 2014), attentional network test (Antón et al., 2014), card sorting test, Simon test and metalinguistic judgments test (Gathercole et al., 2014). In fact, as a recent review by Paap, Johnson and Sawi (2015) points out, the incongruity and inconsistency of the behavioural findings from tasks related to executive functions extend to different paradigms, ranges of ages (i.e., from childhood to the elderly), and types of bilinguals (e.g., early vs. late bilinguals). One could try to account for this discrepancy in the findings by refining the theory in various ways. For instance, the Adaptive Control Hypothesis (Green & Abutalebi, 2013) takes into account the different interactional communicative scenarios and contexts that give rise to varying degrees and types of language switching behaviour, and thus to different specific cognitive and linguistic demands. From this perspective, the specific nature of the bilingual samples from the studies just cited could explain the lack of differences between bilinguals and monolinguals: these studies examine bilingual communities in which dense code-switching between highly interchangeable languages prevails, so these bilinguals do not need to exercise mechanisms of control between their languages in the same way that other bilinguals have to. This refined version of the bilingual advantage hypothesis continues to assume that certain dual-language contexts (e.g. without dense code-switching) give rise to enhanced executive control, and there is evidence to support this hypothesis from both

[1] Note that for the sake of simplicity, we will use the term "bilingual advantage" exclusively to refer to the differences sometimes reported for bilinguals outperforming monolinguals in tasks tapping into executive functions.

children and adults of different ages. However, it should be noted that an increasing number of studies testing bilingual samples from similar dual-language contexts (in which code-switching occurs to a lesser degree) have failed to find differences between bilinguals and controls (Paap & Greenberg, 2013; Paap, Johnson, & Sawi, 2014; Paap & Sawi, 2014). It remains to be seen whether some other factor(s) can explain the lack of a bilingual advantage in these cases. Furthermore, a recent study claimed that a publication bias gives preference to results favouring the so-called 'bilingual advantage' hypothesis (de Bruin, Treccani, & Della Sala, 2015a). While the existence of a publication bias in past years is a controversial issue (see the recent debate between Bialystok et al., 2015, and de Bruin, Treccani, & Della Sala, 2015b), the difficulty in reaching a unified theoretical account given the presence of both null and significant differences between bilinguals and monolinguals in tasks related to different aspects of executive control is uncontroversial.

Hence, despite the large number of studies on this topic, no convergence has been reached on whether bilinguals exhibit better executive functioning than monolinguals at the behavioural level, or on the contexts in which this difference could be observed. One of the main problems for these divergent results could be the scant attention paid to the high variability among language profiles of bilingual individuals (Paap & Greenberg, 2013), which can in turn increase the variability in their ability to control for the interference caused by the non-target language(s) (e.g. Green & Abutalebi, 2013). Another limitation of existing behavioural studies in this regard is the non-systematic use of different tasks that involve very different weights of the components of the executive control system (e.g. monitoring, inhibition, shifting) (Miyake et al., 2000; Friedman et al., 2008; Miyake & Friedman, 2012). Furthermore, as shown by Paap & Greenberg (2013), tasks typically used

to explore some of these components of executive control (e.g. inhibition) do not correlate with each other, pointing to the multidimensional nature of the measures obtained (see also Kroll & Bialystok, 2013).

An important contribution to this debate over behavioural data on bilingualism comes from studies that investigate these issues with neuroimaging methods. Investigating the brain mechanisms underlying these cognitive processes may help to gain a better understanding of the putative bilingual advantage, particularly for identifying which conditions give rise to this cognitive advantage. The biological underpinnings of bilingualism have been approached from both functional and structural perspectives. The search for anatomical changes is almost certainly a necessary preliminary step in this complex task of understanding the specific biological processes underlying bilingualism. Here, we focus on structural work, with some reference to functional connectivity.

The fact that language learning happens so readily – whether it be a child picking up languages effortlessly or an adult, albeit with more effort, learning a foreign tongue late in life in a natural environment or under classroom instruction – points in the direction of neuroplasticity. It is well established that the brain constantly changes structurally and functionally under many challenging situations, and this neuroplasticity plays an important role in learning and memory. Bilingualism, like many other fields of expertise (Carreiras et al., 2009; Draganski et al., 2004; Gaser & Schlaug, 2003; Lee et al., 2007; Maguire et al., 2000), involves structural and functional consequences for the brain. In recent years, a growing body of evidence addressing structural neuroplasticity in bilingualism has begun to emerge (see Li et al., 2014, for a detailed review). However, taken together, the results of these studies demonstrating how brain structure changes due to bilingual experience are

heterogeneous and sometimes conflicting: while some studies have found a variety of neural regions that differ between bilinguals and monolinguals with a certain degree of consistency, others have failed to show any bilingual-specific effect or have reported localized differences in inconsistent brain areas. In contrast to the findings from studies exploring forms of expertise not related to language (Maguire et al., 2000; Maguire, Woollett & Spiers, 2006; Gaser & Schlaug, 2003), the hazy picture obtained from neuroimaging studies of bilingual-specific effects demonstrates that it is unclear where precisely the structural neural differences between monolingual and bilingual samples lie, and what the main factors leading to these structural differences are.

Leaving aside the functional neuroimaging evidence showing a bilingual advantage (see Abutalebi & Green, 2007; Hernandez, 2009; Luk et al., 2012, for a review), which could be influenced by task-related factors boosting multifaceted assessment of the interface between language control and executive control (see Paap & Greenberg, 2013), structural neuroimaging studies seem well-suited to explore task-independent differences between bilinguals and monolinguals in the structure of regions involved in language and general executive control mechanisms. The guiding hypothesis underlying this approach is clear-cut: if it is the case that bilingualism leads to enhanced language-related as well as domain-general executive control processes, then structural differences may be found in the neural regions that underlie these processes. Abutalebi & Green (2007; see Green & Abutalebi, 2013, for an updated and better-defined version) proposed an overall network of regions responsible for cognitive control and bilingual language production. This network is made up of the anterior cingulate cortex (ACC), the left prefrontal cortex (including mainly inferior frontal cortex), the left basal ganglia and the inferior parietal/Supramarginal

gyrus. Abutalebi and Green suggested that a single language network mediates the representation of both languages for a bilingual, and that the executive control network modulates activation of this language network on an adaptive basis depending on the specific characteristics of the language context and the code-switching demands (cf. Green & Abutalebi, 2013). To this end, different studies have explored whether bilingual experience alters the structure of these regions in both grey and white matter, and also in terms of their functioning (see Bialystok et al., 2012; Costa & Sebastián-Gallés, 2014; Li et al., 2014, for reviews). However, as we will detail below, results have been surprisingly inconsistent across studies and show a somewhat erratic pattern of differential effects in these regions as well as the presence of differential effects in other regions outside the proposed network.

The main aim of this present review is to bring together all these recent reports on structural plasticity in bilingualism, which will reveal a great deal of variability in the findings. A second goal will be to draw attention to two important problems that could be underlying this variability in the results: 1) sample selection; and 2) methodological issues. The first problem stems from the heterogeneous nature of the very phenomenon of bilingualism around the globe (Edwards, 2004) resulting in different bilingual profiles for the samples in each study. The variables that contribute to these different profiles include (but are not limited to) the context of learning and acquisition of the L2 (e.g. natural environments vs. artificial environments), the age of acquisition (AoA) of the L2, exposure and/or amount of daily use of the L2, and different proficiency levels of the L2 (see Grosjean, 1998, for an overview). Furthermore, the characterization of a given profile may be skewed by the absence of objective measures to assess language proficiency or the

presence of quantitative measures from only one language (e.g. L2) or from only one aspect of language processing (e.g. semantics). The different combinations of these factors may potentially produce results that are not comparable.

In this review, we compare structural studies performed with children, younger adults and older adults. These studies investigated simultaneous, sequential or/and late bilinguals. Comparing these different groups and bilingual profiles aids in appreciating how brain changes are conditioned by age ranges and different AoAs for the L2. Even so, the small number of studies severely limits the conclusions. There has been much discussion of the critical period for learning a second language; as with learning other non-linguistic skills, the learning of an L2 is undeniably affected by the age at which learning begins (Hernandez and Li, 2007). Many alternative hypotheses have emerged concerning these effects in bilingualism (see Hernandez, Li & MacWhinney, 2005; Hernandez & Li, 2007, for an overview). The precise moment that L2 learning begins during development will determine which domains are more sensitive to learning. In general, AoA effects are based on developmental constraints, especially the maturation of sensorimotor processing (see 'The sensorimotor hypothesis', Hernandez & Li, 2007). In order to take into account how structural changes evolve across time, we also discuss correlational studies between AoA and L2 proficiency with brain structural measures.

The second problem concerns the lack of consistency in the methods used for the analysis of the brain, such as voxel-based morphometry (VBM) and/or region of interest (ROI) analysis. (See Appendix 1 for details.) VBM is an automated whole-brain (i.e. including every voxel of the brain) magnetic resonance image (MRI) measurement technique, whereas ROI-based measurement typically involves manual delineation or

automatic extraction of the ROIs and the averaging of the MRI signal extracted from the voxels included in the ROI. There has been much debate about the use of voxel-based or ROI-based approaches (Good et al., 2001), and the relevant issues will be discussed throughout this article. Additionally, there are various criticisms of VBM (i.e. preprocessing steps: segmentation, registration algorithms, modulation of the images after registration, etc.) that could be limiting the bilingualism studies. Finally, as with any field of research using MRI, it is important to consider the uniformity of the data used for the analysis of the brain.

Our aim is to discuss the two major methodological topics mentioned above and subsequently to summarize current findings. In addition, we will examine how evidence for structural changes may contribute to this debate on the 'advantages' and 'disadvantages' of bilingualism at the behavioural level (Abutalebi & Green, 2007; Antón et al., 2014; Bialystok & Barac, 2012; Costa et al., 2009; Costa et al., 2008; Duñabeitia et al., 2014; Gathercole et al., 2014; Gollan et al., 2011; Kroll & Bialystok, 2013; Martin et al., 2012; Paap & Greenberg, 2013; Paap et al., 2014; Paap & Sawi, 2014). In this sense, we will highlight how neuroimaging data could contribute to the debate by adding empirical evidence from a different perspective. Finally, we conclude by identifying issues that should be taken into account so that studies in this field are more comparable by providing evidence that could be collected, processed and integrated more easily.

This review is organized as follows. Section 2 discusses cross-sectional studies of structural brain changes in bilingualism and is divided into grey matter (GM, section 2.1) and white matter (WM, section 2.2) studies since evidence shows that they can vary independently (Li et al., 2014). After dealing with the group comparisons for both GM and

WM, there is also a separate subsection (2.3) looking specifically at the evidence from correlation analysis showing the effects of AoA and L2 proficiency on the brain. Such correlational studies offer valuable insight into what factors may drive structural changes in the brain in the context of bilingualism and provide a better understanding of how these changes evolve across time. Nevertheless, the results described in sections 2.1 and 2.2 based on group comparisons are more robust and provide direct evidence that shows how the brain changes in bilingualism. Importantly, in order to make the conclusions as clear as possible, this review only takes into consideration results that have been corrected for multiple comparisons, since uncorrected results just show a tendency and cannot be generalized. Section 3 reviews longitudinal brain studies of short-term immersion learning or intensive training in the L2. In contrast to cross-sectional studies that investigate long-term bilinguals and can thus show more stable changes in the brain, longitudinal studies show transient changes related to the process of learning and briefly (but intensely) experiencing a second language. Section 4 is dedicated to functional and structural brain connectivity studies conducted in this field and provides a brief insight into how functional/structural connectivity may contribute to the debate. The final section summarizes the most specific brain changes in bilinguals described in this review and discusses the main methodological differences among studies that bring about so many inconsistencies in the field. The review closes with methodological recommendations to follow in future studies with the aim of providing a methodological framework that will help the field to progress.

2. Cross-sectional studies of structural brain changes in bilingualism

2.1. Grey matter studies

For the study of GM researchers have typically used high-resolution T1-weighted MRI to obtain measures such as grey matter volume or density and cortical thickness (CT) of the brain (see Appendix 1 for a description of each measure). Measuring volume/density of the grey matter typically involves VBM and ROI analysis. The thickness of the cerebral cortex is also another measure that can be automatically extracted from the T1-weighted MRI, which allows cross-subject statistical comparisons to be performed in order to detect focal changes in the brain (Fischl & Dale, 2000). (See Appendix 1 for a description of each technique.)

2.1.1. Volume/density studies

2.1.1.1. Using whole-brain approach

Mechelli et al. (2004) was the seminal study indicating anatomical changes in the brain for bilinguals as compared to monolinguals. They compared 25 early English-Italian bilinguals (who started to learn their L2 before the age of 5), 33 late bilinguals (who started to learn the L2 between 10 and 15 years old), and 25 English monolinguals. All groups were comparable in age and educational level. VBM analysis of the GM density, using the statistic parametric mapping (SPM) software package (available at http://www.fil.ion.ucl.ac.uk/spm/), revealed significant GM increases for the bilinguals in the left inferior parietal lobule (IPL) corrected for family wise error (FWE) at voxel-level (see Figure 1a, red and Table 1).

More recent studies have also obtained significant difference effects between bilinguals and monolinguals using different methods. Pliatsikas, Johnstone & Marinis (2014) compared 17 Greek-English bilinguals (mean age, 27.5; mean L2 AoA, 7.7; mastery proficiency in the L2) with 22 English monolinguals (mean age, 24.5). They performed a whole brain comparison using the threshold free cluster enhancement (TFCE) technique (Smith & Nichols, 2009) implemented in the FSL software (Smith et al., 2004) to correct the FWE. This showed a large increment of GM volume for bilinguals in the cerebellum (Pliatsikas et al., 2014) (see Figure 1a, blue and Table 1). However, due to VBM limitations in this area related to poorer segmentation (Ashburner & Friston, 2000), Pliatsikas et al.'s results require replication with a different set of subjects.

Abutalebi et al. (2014) performed a VBM study using SPM, comparing 23 older adult bilinguals (12 Cantonese-English and 11 Cantonese-Mandarin; mean age, 62.2; mean L2 AoA, 18.87) with 23 Italian monolinguals (mean age, 61.9). This study obtained a significant volume increase for bilinguals in the left anterior inferior temporal gyrus (aITG) (see Figure 1a, yellow and Table 1) using cluster-level correction of the FWE in SPM, a different method of inference from previous studies.

Abutalebi, Guidi, Borsa, Canini, Della Rosa, Parris & Weekes (2015) performed a VBM study using SPM, comparing 19 older adult bilinguals (11 Cantonese-English and 8 Cantonese-Mandarin; mean age, 61.68; mean L2 AoA, 12.68) with 19 Italian monolinguals (mean age, 60.93). The results showed a significant volume increase for bilinguals in the left/right ACC (see Figure 1a, green and Table 1) using FWE cluster-level correction.

Despite these results, other studies also performing VBM analysis failed to find significant differences (see Table 1) between bilinguals and monolinguals correcting for multiple comparisons across the whole brain (Gold, Johnson, & Powell, 2013; Grogan et al., 2012; Ressel et al., 2012). Gold et al. (2013) studied 20 older adult English native bilinguals who started to learn the L2 before the age of 5; the L2 was variable between participants and the mean age of the group was 63.9 years old. The study compared the bilingual group with 20 English monolinguals with a mean age of 64.4 years old. Grogan et al. (2012) studied 31 young multilingual adults who learned English as L2 (the native and other languages varied between participants); the mean age of the group was 30.9 years old. They compared the multilingual group with 30 young non-native English bilingual adults (who also had different native languages); the mean age of the group was 30.6 years old. Although the L2 AoA was variable in both the multilingual and bilingual groups, it was balanced between them. Notice that the sample profiles are very different between these two studies and also different from the studies described above. In contrast, the study by Ressel et al. (2012) used a more similar sample profile as the Mechelli et al. (2004) study. Even so, this study also failed to find any significant differences at voxel-level. They compared 22 young Catalan-Spanish bilinguals who started to learn the L2 before the age of seven (mean age, 23.1) and 22 Spanish monolinguals (mean age, 21.5).

In summary, these studies explicitly looked at whether or not the bilingual brain differs from that of the monolingual, correcting the FWE across the whole brain. On the one hand, differences appear exclusively in three regions: the ***left IPL*** (Mechelli et al., 2004), the ***cerebellum*** (Pliatsikas et al., 2014), the ***left aITG*** (Abutalebi et al., 2014) and the ***ACC*** (Abutalebi, Guidi, et al., 2015) (Figure 1a, Table 1). However, different levels of

inferences (i.e. voxel-based or cluster-based) were used, which means different levels of sensitivity: cluster-level inferences are more powerful than voxel-level inferences but also imply less localizationist power. The studies also used different FWE controlling methods (i.e. random field theory (RFT) or TFCE and permutations). Consequently, more studies are needed to confirm these results. Conversely, there are three studies that consistently showed negative results: no differences between bilinguals and monolinguals (Gold et al., 2013; Grogan et al., 2012; Ressel et al., 2012) (Table 1). So far, only one study has found a bilingualism effect in an expected region: the left IPL. Nevertheless, as the next section will show, when these studies limited their analysis to a region or volume of interest, effects start to appear in expected regions. In any case, these negative results provide interesting findings and help researchers in the field to form new hypotheses.

[Figure 1 near here]

2.1.1.2. Using ROI-based approach

Some of the studies described in the previous section also used a ROI approach to compare groups. For example, Ressel et al. (2012) manually extracted the mean volume from the right and left Heschl gyri to compare between bilinguals and monolinguals, and obtained significantly larger volumes in bilinguals, bilaterally. Using the automatic anatomical labelling atlas (Tzourio-Mazoyer et al., 2002), Abutalebi et al. (2014) extracted the mean volume for the right/left temporal pole (TmP) and right/left orbito-frontal cortex (OFC), and found greater mean volume for bilinguals as compared to monolinguals in both regions and hemispheres.

In a more recent study, Abutalebi, Canini, Della Rosa, Green & Weekes (2015) extracted GM volume from ROIs in the left/right IPL using the coordinates reported by Mechelli et al. (2004). They studied 30 older bilinguals (mean age, 63.2; 16 Cantonese-English bilinguals and 14 Cantonese-Mandarin bilinguals) who started to learn the L2 at a mean age of 18.27 years, and compared them to 30 older Italian monolinguals (mean age, 61.85). They found that the volume in the left/right IPL was significantly greater for the bilingual group.

Olsen et al. (2015) extracted the volume of the GM for the frontal, temporal, parietal and occipital lobes for both right and left hemispheres. They investigated structural differences in the brain of 14 older bilingual adults (mean age, 70.4) who reported regular use of both English and another alphabetic language since before the age of 11. They compared the bilinguals with 14 English monolinguals (mean age, 70.6) and did not obtain any significant group effects in their GM analysis.

Although ROI analysis increases statistical power with respect to whole-brain analysis, the use of ROIs can limit the fine-grain spatial resolution of the effect of interest. Additionally, this type of analysis can miss true differences as a result of the averaging if the variation in the entire ROI is not uniform because parts with no significant difference and parts with a significant difference may be averaged over in the same ROI. The result is that a significant effect is blurred over an entire region. Or, conversely, such averaging may highlight differences if there is a fairly uniform difference that is not very great, rendering a significant effect that would not be deemed significant in a voxel-based analysis after correction for multiple comparisons.

Alternatively, some studies performing VBM used small volume corrections (SVC) as a way of limiting the analysis to specific regions without suffering from the problems of ROI averaging. This can be helpful as a middle point between ROI and whole-brain approaches. Two of the previously mentioned studies used this approach after failing to find differences when correcting across the whole brain. Grogan et al. (2012) found an effect not in the left but in the right IPL using Mechelli et al.'s coordinates for the SVC. However, the comparison was between multilinguals vs. bilinguals instead of bilinguals vs. monolinguals (see Figure 1b, green and Table 1). In their VBM analysis, Ressel at al. (2012) additionally performed SVC on the Heschl gyri and differences appeared in just the left hemisphere (Figure 1b, red and Table 1).

Two other studies have used SVC implemented in the SPM software but without reporting whole-brain results. Abutalebi et al. (2013) studied 14 German-Italian-English multilinguals who started to learn the L2 before the age of five and the third language (L3) after the age of ten, comparing them with 14 Italian monolinguals. The mean age for both groups was 23.5 years old. After SVC they obtained higher GM volume in the left putamen for multilinguals as compared to monolinguals (Figure 1b, blue and Table 1). However, this was done using false discovery rate (FDR), which is a different correction to the FWE. In addition, Zou et al. (2012) studied 14 bimodal Chinese/Chinese Sign Language (CSL) adult bilinguals, (mean age, 49; mean L2 AoA, 19; 29 years of experience with CSL). They compared the bimodal Chinese-CSL bilinguals with 13 Chinese monolinguals (mean age, 48) and found an increased volume after SVC for bilinguals in the left caudate (Figure 1b, purple and Table 1).

In summary, when some of the studies performing VBM limited their analysis to the scope of certain regions of interest effects started to appear in the ***Heschl gyri*** (Ressel et al., 2012), the ***right/left TmP*** and ***OFC*** (Abutalebi et al., 2014), the ***right IPL*** (Grogan et al., 2012; Abutalebi, Canini et al., 2015) the ***left putamen*** (Abutalebi et al., 2013) and the ***left caudate*** (Zou et al., 2012). These are isolated results, with the exception of the ***right IPL***. Unfortunately, there is no uniformity in the samples compared across these studies: for example, bilinguals vs. monolinguals (Ressel et al., 2012), multilinguals vs. bilinguals (Grogan et al., 2012), multilinguals vs. monolinguals (Abutalebi et al., 2013), bimodal bilinguals vs. unimodal monolinguals (Zou et al., 2012). Therefore, the origin of these effects is variable and may not represent a clear effect of bilingualism. In addition, it is striking that only two studies investigated the same region (namely, the IPL), making it extremely difficult to arrive at any solid conclusions from these ROI analysis. Again, differences between bilinguals/multilinguals and monolinguals have been found in different regions in different studies. Equally important, the quality/noise of the data is variable across these studies since some of them used 1.5T and others 3.0T MR scanners (see table 1).

[Table 1 near here]

2.1.2. Cortical thickness cross-sectional studies

Klein et al. (2014) performed a cross-sectional cortical thickness study on 12 simultaneous bilinguals (mean age, 23; AoA, below 3 years old), 25 early bilinguals (mean age, 26; L2 AoA, after 4 years old and before 7 years old; mean L2 AoA, 5 years old), and 29 late bilinguals (mean age, 28; L2 AoA, after 8 years old and before 13 years old; mean L2 AoA, 10 years old), all French-English bilinguals. They compared the bilingual groups

with a control group of 22 English monolinguals (mean age, 25). The results showed greater cortical thickness for early and late bilinguals as compared to monolinguals in the ***pars triangularis (IFGTr)*** and ***pars orbitalis (IFGOr)*** of the ***left IFG (inferior frontal gyrus)*** and less cortical thickness in the ***right IFGOr*** for late bilinguals as compared to monolinguals.

Olsen et al. (2015) performed a CT analysis on their bilingual and monolingual samples (described above, see section 2.1.1.2) in the entorhinal cortex and temporal pole, but they did not find any group differences. What they observed was a significant negative correlation between the thickness of the temporal pole and age in the monolinguals but not for bilinguals.

To date these are the only cross-sectional studies investigating cortical thickness. Interestingly, again these results do not replicate the GM volume/density results summarized above.

2.2. White matter studies

Recently, there have been an increasing number of studies investigating WM changes related to bilingualism. The first study looking for WM differences between bilinguals and monolinguals was Mechelli et al. (2004), who, in addition to the GM analysis, used a VBM analysis to look for differences in WM volume but failed to detect any differences across the whole brain (see Table 2). Ressel et al. (2012) also looked for WM differences using the same approach and found no differences either (see Table 2). However, most studies looking at WM changes have employed diffusion tensor imaging (DTI) (see Appendix 1) instead of T1-weighted MRI.

Almost all the studies employing DTI have employed tract-based spatial statistics (TBSS, see Appendix 1 for a description) (Smith et al., 2006), implemented in the FMRIB software library (FSL) (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012), to compare images using different diffusion measures: mean diffusivity (MD), axial diffusivity (AD), radial diffusivity (RD), and fractional anisotropy (FA). (See Appendix 1 for explanations of these measures.) TBSS protocol typically used Threshold free cluster enhancement (TFCE) and non-parametric permutation testing to reach statistical inferences and to control the FWE rate, also implemented in FSL.

Luk et al. (2011) studied 14 older adult native English bilinguals who started to learn a L2 after the age of 11 (the L2 was variable across participants) and compared them with 14 English monolinguals. The mean age between groups was 70.5 years old. They performed a TBSS analysis and found significantly increased FA values for bilinguals as compared to monolinguals in parts of the corpus callosum (CC) (Figure 2, green) that extended bilaterally into Superior Longitudinal Fasciculus (SLF) (Figure 2, blue) and into to the right Inferior Fronto-Occipital Fasciculus (IFOF) (Figure 2, red) and Uncinate. They also obtained significantly decreased RD values for bilinguals in the body of the CC, overlapping with some of the areas of increased FA (Luk et al., 2011) (see Table 2).

Gold et al. (2013) obtained a different result using the same approach to study 20 English bilinguals (mean age, 63.9) who started to learn the L2 after the age of five (the L2 varied across participants). They matched the bilinguals with 20 English monolinguals (mean age, 64.4). The authors obtained a significant decrease in the FA values for bilinguals as compared to monolinguals in many portions of the CC (Figure 2, green) and bilaterally in the inferior longitudinal fasciculi (ILF), IFOF (Figure 2, red) and fornix. They

also obtained significantly increased RD values in regions of reduced FA, particularly in IFOF and CC but also in smaller parietal and occipital tracts. These results do not support those of Luk et al. (2011). This difference may be due to the fact that the samples in each study were slightly different (see Table 2).

Cummine & Boliek (2013) studied young adult Chinese-English bilinguals (mean age, 24.2; L2 AoA before the age of five) and 11 English monolinguals (mean age, 28.5). They obtained significant decreases of the FA for bilinguals as compared to monolinguals in the right IFOF (see Figure 2, red and Table 2), which is in line with Gold et al. (2013). They also obtained decreased FA in the anterior thalamic radiation, especially in the right superior portion and bilaterally in the inferior portion.

Recently, Pliatsikas, Moschopoulou & Saddy (2015) studied 20 sequential bilinguals (mean age, 31.85) who had highly proficient English as L2 (mean AoA, 10.15; mean immersion, 91 months; L1 varied across participants) and were highly proficient in English. The bilinguals were compared with 25 English monolinguals (mean age, 28.16). The authors performed a TBSS analysis, revealing higher FA values for these sequential bilinguals bilaterally in the whole CC (genu, body and splenium), the IFOF, the uncinate and the SLF (see Figure 2 and Table 2). These findings are in line with Luk et al. (2011).

[Figure 2 near here]

A different approach for the analysis was taken in the study by Mohades et al. (2012), who used a tract of interest (TOI) approach to compare FA maps instead of the TBSS approach. Their method involved first reconstructing the fibre tracts for the whole brain using a tractography algorithm. Then manually defined ROIs were used to separate 4

TOIs. Finally, the FA values from the voxels included in these tracts provided a mean FA value for each tract and each individual. The samples consisted of 15 simultaneous bilingual children who started to learn the L2 before the age of three (mean age, 9.3) and 15 sequential bilinguals who started to learn the L2 after the age of three (mean age, 9.7). The native language of all bilingual children was Dutch and the L2 was variable between participants. The control group was 10 Dutch monolinguals (mean age, 9.6). The authors obtained higher mean FA values for the bilinguals as compared to the monolinguals in the IFOF, and lower mean FA values in the tracts going from the anterior part of the CC to the orbital lobe (see Table 2). The higher mean FA value for the IFOF in these bilingual children (Mohades et al., 2012) is the opposite pattern found for young adult bilinguals in Cummine & Boliek (2013), but the same pattern as Pliatsikas et al. (2015) and also for older adult bilinguals (Luk et al., 2011). At the same time, the lower mean FA value in the CC obtained for these bilingual children is opposite to the pattern obtained for older adult bilinguals in the Luk et al. (2011) and Pliatsikas et al. (2015) studies but in line with the pattern obtained for older bilinguals in the Gold et al. (2013) study.

Olsen et al. (2015) also extracted the volume of the WM for the frontal, temporal, parietal and occipital lobes in both hemispheres of their sample (described above, see section 2.1.1.2). They found that the WM volume in the frontal lobe was significantly higher for their older adult bilingual sample as compared to monolinguals (see Table 2).

In summary, two WM regions seem to be the focus of neuroplasticity in bilingualism, namely the ***CC*** and ***IFOF***. However, while some studies found increased FA values in the ***CC*** for older and younger adult bilinguals (Luk et al., 2011; Pliatsikas et al., 2015) others found decreased FA values for older adult bilinguals (Gold et al., 2013) and

children (Mohades et al., 2012). And while some studies found increased FA in the ***IFOF*** for older adult bilinguals (Luk et al., 2011), younger adults (Pliatsikas et al., 2015) and children (Mohades et al., 2012), others found decreased FA values for older (Gold et al., 2013) and younger (Cummine & Boliek, 2013) adult bilinguals. Since almost all of the WM studies use the same methodological approach (i.e. TBSS) in the analysis of the diffusion-derived measures, this makes them more comparable to each other than the GM studies are, and yet they show many inconsistencies and sometimes the results are completely contradictory. However, there are several confounding factors among the samples, such as the chronological mean age and the age of acquisition of the L2. Consequently, are these inconsistencies due to a combination of maturation/degeneration processes and second language acquisition processes? Previous studies have demonstrated that WM declines linearly with age in some local areas, such as CC, internal capsule and prefrontal regions, while other areas remain relatively preserved, such as temporal and posterior regions (Good et al., 2001; Salat et al., 2005). As such, the two focal regions for bilingualism (CC and IFOF) seem to be particularly vulnerable to age effects, and more studies are needed in order to clarify these findings. Thus, greater interaction with work on changes in WM during development and aging is required to be able to progress in this area. Importantly, the quality/noise of the data is variable across these studies since some of them used 1.5T and others 3.0T MR scanners and the acquisition parameters (such as number of directions and voxel size) of the DW-MRI vary across studies (see table 2).

[Table 2 near here]

2.3. The effect on the brain of age of acquisition and proficiency in L2

Some cross-sectional studies mentioned in previous sections have also investigated the effect of AoA and L2 proficiency on the brain. Although this is not the main focus of this review it deserves mention on two counts. Firstly, considering the brain as a non-linear dynamic system and (individual) bilingualism as a dynamic process (Hernandez, 2013), it is relevant to consider both the point at which bilingualism begins to influence the system (i.e. AoA) and how the effect of bilingualism cumulatively interacts with the system (i.e. L2 proficiency and AoA). Secondly, adaptive models of the neural underpinnings of bilingualism (e.g. Green & Abutalebi, 2013; Abutalebi & Green, 2007) take into account the dynamic development of the brain networks in question over time. Therefore, this section looks at the relationship between second language experience and brain structure.

The study by Mechelli et al. (2004) used SVC around the region where they obtained the group effect – the left IPL – and obtained a negative correlation between GM density and L2 AoA, which means that GM density increased as the L2 AoA decreased. Additionally, they obtained a positive correlation between the GM density and L2 proficiency, which means that GM density increased as proficiency increased. Grogan et al. (2012) also showed a positive correlation between GM volume and L2 proficiency in the pars opercularis (IFGOp) of the left IFG.

Similarly, Hosoda et al. (2013) found the same positive correlation in the IFGOp and also in FA values of WM tracts beneath the right IFGOp and inside the right ILF and Arcuate fasciculus, two tracts that typically connect language regions. They studied 137 Japanese-English bilinguals who started to learn the L2 after the age of 7 (mean age, 11). In addition, they performed a tractography analysis and also found a positive correlation

between L2 proficiency and connectivity in the right pathways connecting IFGOp to caudate, and IFGOp to superior temporal gyrus (STG)/supramarginal gyrus (SMG).

Using a ROI analysis, Abutalebi et al. (2013) showed that the GM volume in the left putamen increased as proficiency in the third language increased, and they reported this effect solely for the third language because no correlation effects were obtained for either the native language or the L2. In a later study also using a ROI approach, Abutalebi et al. (2014) found a positive correlation between GM volume in the left TmP and proficiency in L2 in a group of multilingual subjects. Additionally, the Abutalebi, Canini et al. (2015) study found no significant correlation between the L2 AoA and the IPL GM volume, but did reveal a positive correlation between the L2 naming performance and the GM volume in the left IPL, and between the L2 exposure time and the GM volume in the right IPL.

On the other hand, Klein et al. (2014) showed that CT correlates positively with L2 AoA in the left IFG and the left superior parietal lobule (SPL) for bilinguals, and negatively in the right IFG. It is important to note that CT is a different measure and has a different interpretation to that of GM volume. Differences in local GM volume can arise from differences in CT and variation in surface area due to the folding pattern (Kanai & Rees, 2011). However, GM volume is more correlated with surface area and much less correlated with CT (Winkler et al., 2010). There are studies (Chung, Dalton, Shen, Evans, & Davidson, 2007) showing a negative correlation between CT and GM density. Thus, although CT results are difficult to interpret, it seems that Klein et al.'s correlation results (Klein et al., 2014) between CT and L2 AoA are in line with correlation results in GM volume/density and ultimately are also consistent with what would be expected for the IFG and parietal regions (Kanai & Rees, 2011; Winkler et al., 2010).

In conclusion, the most consistent effect regarding AoA and L2 proficiency in the brain is in the GM of the ***IFG***, and the WM connecting the ***IFG*** with other GM regions, such as ***caudate, STG*** and ***SMG/IPL***. These results suggest that mastery of the L2 (i.e. increased L2 proficiency and L2 experience) is associated with higher GM volume, higher WM connections and less CT in regions related to executive functioning, specifically the IFG. Abutalebi and Green's model (Abutalebi & Green, 2007; Green & Abutalebi, 2013) predicts that the degree of involvement or activation of these regions changes as a function of L2 proficiency such that there is less involvement when the L2 is mastered and automatized. Bringing together the predictions of the model and the findings for structural changes in the brain described here requires a clear explication of the relation between function and structure, particularly of how 'involvement' and 'activation' spell out in terms of structure. Very tentatively, if (as Abutelebi and Green predict) greater L2 proficiency involves more automatic processing of the language, greater ability for control and thus less activation of the associated regions in the control network, these results suggest that these changes are associated with greater GM volumes and WM connectivity in these regions. This is mere speculation and only further testing can shed more light on the issue.

3. Longitudinal studies of structural brain changes in L2 learners

Most of the studies investigating structural brain changes related to bilingualism are cross-sectional studies that have focused on bilinguals who have already learned and experienced the L2 for long periods of time. Nevertheless, there are other studies looking at how the brain changes during the process of learning a L2. Four longitudinal studies examine short-term immersion learning or intensive training in the L2; three of them straightforwardly address the question of differences between L2 learners and non-learners.

Additionally, another longitudinal study has looked at L2 learning in children over a longer time scale.

Schlegel et al. (2012) investigated WM changes during 9 months of intensive Chinese learning without immersion. They studied a training group of 11 English monolingual learners and a control group of 16 English monolingual non-learners. The mean age for both groups was 20 years old. They obtained the most significant FA increase in the genu of the CC, corrected across the whole brain. They also found increased FA and decreased RD in left frontal language-related regions and in the counterpart regions on the right hemisphere. For the whole-brain analysis, they used a non-parametric permutation test and TFCE to achieve significant cluster effects. Interestingly, since they acquired (nine) monthly MR images from participants, they were able to show that the global mean FA (extracted from all the voxels that showed significantly increased FA in the prior whole-brain analysis) described a significant linear increase over the nine time points for learners. They also showed that the amount of increased FA correlated positively with the amount of language learned across these time points. Additionally, they extracted mean FA and RD values from 111 TOIs that showed higher connectivity between language regions and found increased FA and decreased RD mean values for learners as compared to non-learners in 16 of these TOIs: 5 of them terminated in the caudate nucleus and 10 of them connected together different frontal regions of the left hemisphere or frontal regions between hemispheres (these connections passed through the genu of CC). Even though the sample is small, this is a very germane experimental design since it allows for variability between brains to be eliminated. If changes related to bilingualism are small (and this is a very

plausible scenario), this strategy of analysis is more beneficial than the cross-sectional studies described above.

Mårtensson et al. (2012) performed a vertex-wise CT study comparing 14 native Swedish interpreter students (mean age, 20) who took a 3-month intensive language course focusing on vocabulary for different languages (4 Arabic, 8 Dari and 2 Russian) and 17 native Swedish non-learners (mean age, 21) as a control group. The learner group showed increased cortical thickness in left dorsal middle frontal gyrus (MFG), IFG and STG. Volume measures from left and right hippocampus (the volumetric study was restricted to these regions) revealed larger volume on the right side for learners as compared to non-learners. Additionally, the CT in left STG and volume in right hippocampus correlated positively with proficiency in the L2.

Hosoda et al. (2013) studied Japanese students of English, 24 of whom received 4 months of laboratory training on vocabulary and 20 of whom did not. Mean age for both groups was 20 years old. The results of VBM on GM segmentations and TBSS on FA maps showed a training by group interaction effect in the IFGOp (i.e. increased GM volume and FA values for learners as compared to controls after training). They also traced 8 pathways known to be related to language: IFGOp-to-Caudate, IFGOp-to-STG (dorsal language pathway), IFGTr of the IFG-to-MTG (ventral language pathway) and the ILF in each hemisphere. The results showed increased connectivity for the right IFGOp-to-caudate and IFGOp-to-STG pathways. There was also a positive correlation between the gain in L2 proficiency and the connectivity values on the IFGOp-to-caudate pathway.

The study by Stein et al. (2012) did not perform group comparisons but just described GM changes based on a correlation analysis with behavioural measures. Specifically, they studied GM density changes in a group of 10 English monolinguals after a brief immersion in a L2 (5 months of learning German in Switzerland), showing a positive correlation between the increase of L2 proficiency and the increase of the density in the left IFG and also in the anterior temporal lobe (aTL), using cluster-level correction of the FWE. Notice that this is the only VBM longitudinal study of L2 learning within an immersion context.

Another longitudinal study has looked at language acquisition over a longer time period in simultaneous and sequential bilingual children (with monolingual controls). Mohades et al. (2015) carried out a follow up study on the same 40 children previously tested (Mohades et al., 2011, described in section 2.2). In the first study (time 1) the children had a mean age of 9 years old and in this second assessment (time 2) they were two years older. The authors used the TOI approach described above (in section 2.2), limiting the investigation to those tracks related to language processing. They obtained the mean FA for four language-related pathways and for one bundle as a control: IFOF, SLF, bundle from the anterior part of the CC to orbital lobe, fibre from anterior-midbody CC to motor cortices and right IFOF (as the control pathway not related to language). At time 1 there was a group effect showing higher mean FA values in the left IFOF between bilinguals and monolinguals (see section 2.2). At time 2 the results revealed a group by time interaction for the same bundle. Simultaneous bilinguals had the highest mean FA value in the left IFOF as compared to sequential bilinguals and monolinguals at time 2. Interestingly, the lower mean FA value they observed at time 1 in bilinguals in the anterior

part of the CC (see section 2.2) was no longer evident at the later observation. This study showed interesting results and more studies in children would provide greater insight into the structural evolution that accompanies second language learning.

In summary, these studies provide evidence that left/right frontal regions, especially the ***IFG, MFG***, but also the ***STG*** and ***caudate***, as well as the WM connecting all these regions, especially the ***CC*** connecting frontal regions and pathways connecting frontal regions with caudate, are the targets for plastic changes when a L2 is acquired and improved. In general, longitudinal studies show more consistent results than the cross-sectional studies described above. Since longitudinal studies investigate brain changes within the same subject across time under different conditions, they may represent a much better option than cross-sectional designs to avoid inter-subject variability. However, there is no denying that this kind of experiment is more expensive and time consuming.

4. Brain connectivity studies in bilingualism

Studies of brain networks examine the relationship and interaction between brain regions to provide more complete information about the organization and configuration of these regions and the brain as a whole. Potentially, this might offer a better understanding of the possible mechanisms underlying the cognitive processes associated with learning and using a second language. This section describes the few studies that have investigated both structural and functional relationships between GM regions in bilingualism in order to provide an account of brain connectivity.

Luk et al. (2011) combined WM results from the TBSS analysis (see section 2.2) with resting-state (RS) functional connectivity measures. They performed a RS functional

connectivity analysis taking as seeds the regions of GM adjacent to the cluster showing higher FA values for bilinguals in the prior TBSS analysis and which they considered important for language switching (i.e. right/left IFG). This analysis showed stronger functional connectivity between left IFG and posterior brain regions (i.e. with middle temporal gyri, right IPL, precuneus, middle occipital gyri and left caudate) for bilinguals as compared to monolinguals. In contrast, monolinguals showed a different connectivity pattern, showing higher connectivity between left IFG and other frontal regions.

García-Pentón et al. (2014) investigated WM structural brain connectivity differences between 13 early Basque–Spanish bilinguals (mean age, 24.08; L2 AoA before 3 years old) and 13 Spanish monolinguals (mean age, 29.07). They performed an anatomical connectivity analysis and complex network analyses based on DW-MRI. The connection density between pairs of GM regions was estimated from a tractography algorithm. A network-based statistic (Zalesky, Fornito, & Bullmore, 2010) approach and complex network analysis was employed to identify differences in connectivity patterns and properties of the networks between both groups. The authors identified two different sets of regions (subnetwork I and subnetwork II) interconnected by anatomical tracts that were more strongly connected and graph-efficient in early bilinguals as compared to monolinguals. Sub-network I contained left frontal and parieto-temporal brain regions, most of them previously described in the literature as language-relevant: Insula—STG—IFGTr—SMG—IFGOp—Medial Superior Frontal Gyrus. Sub-network II also included some brain regions that have also been extensively related to language processing (i.e. left angular gyrus (AnG) and left superior TmP) while the others have been implicated in other

cognitive processes related to language: left Superior Occipital gyrus—right Superior Frontal Gyrus—left SPL—left superior TmP—left AnG.

In summary, both functional and structural connectivity studies consistently identified the ***left IFG***, a region related to cognitive/language control, and showed how this region is related to a more extended set of regions. These studies are particularly useful for studying large-scale structural and functional connectivity plasticity associated with many cognitive functions (Guye, Bartolomei & Ranjeva, 2008), such as language and executive functioning, a topic that is largely unexplored. The search for differences between bilinguals and monolinguals cannot be limited to locating different structures but must also look at patterns of functional and structural connectivity. If there are bilingualism effects, they may be evident not as a change in the volume of a region, but as the connections between the different regions of a circuit.

5. Conclusions

To conclude, nine studies performing VBM looked for GM differences between bilinguals and monolinguals across the whole brain. Three of them were cross-sectional studies that reported no significant brain differences (Gold et al., 2013; Grogan et al., 2012; Ressel et al., 2012). In contrast, three other studies using different techniques/measures and experimental designs consistently reported GM changes in the IFG: Klein et al. (2014) obtained increased cortical thickness for bilinguals in the left IFG but decreased cortical thickness in the right; Mårtensson et al. (2012) also found increased CT in the left IFG, and Hosoda et al. (2013) obtained increased GM volume in the right IFG. These last two studies looked at intensive L2 learning experiences longitudinally. Finally, three cross-sectional studies performing VBM each found increased density/volume in a different region: the left

IPL (Mechelli et al., 2004), cerebellum (Pliatsikas et al., 2014), left aITG (Abutalebi et al., 2014) and ACC (Abutalebi, Guidi et al., 2015). Each of the studies used different methods for the preprocessing and analysis of the data (see table 1).

Eight studies performed ROI analysis, extracting mean volumes from the regions or reducing the analysis to the scope of a volume of interest. Two studies showed increased GM in the right IPL (Grogan et al., 2012; Abutalebi et al., 2015), and the rest showed isolated results.

Considering WM, four cross-sectional studies looked for differences between bilinguals and monolinguals across the whole brain using TBSS. The most consistent WM changes were observed in CC and IFOF. However, while two studies found increased FA values in CC for bilinguals (Luk et al., 2011; Pliatsikas et al., 2015), another found decreased FA values (Gold et al., 2013). And some studies found increased FA values in IFOF for bilinguals (Luk et al., 2011; Pliatsikas et al., 2015), others found decreased FA values (Cummine & Boliek, 2013; Gold et al., 2013). On the other hand, two other studies performing VBM reported no significant differences in WM volume (Mechelli et al., 2004; Ressel et al., 2012). However, these last two studies used a completely different measure and methodology to those of the former studies. In line with results in cross-sectional studies, there are also two longitudinal studies: one showed increased FA in CC for learners vs. non-learners (Schlegel et al., 2012) and another one showed increased WM volume inside the right IFG (Hosoda et al., 2013).

Regarding brain network connectivity studies, Luk et al. (2011) showed that the left IFG in bilinguals as compared to monolinguals had stronger functional connectivity with

posterior brain regions in the temporal, parietal and occipital gyri but had a different functional connectivity pattern in monolinguals, who showed higher connectivity between the left IFG and other frontal regions. García-Pentón et al. (2014) revealed that early bilinguals showed a different WM structural configuration of the brain, developing more highly interconnected and efficient subnetworks to achieve the processing of the two languages, and that these changes seem to be at the expense of decreased efficiency for the whole brain network. This is in line with previous accounts broadly showing that bilinguals are less accurate and slower than monolinguals of each language in linguistic tasks (e.g. picture naming, word recognition, lexical decision) (Gollan et al., 2011; Martin et al., 2012), under the supposition that the over-developed structural subnetworks allow bilinguals to deal with two languages but do not improve linguistic skills per se in each language. Furthermore, Luk et al. (2011) showed stronger functional connectivity between the left IFG and other frontal regions in monolinguals and this pattern could be important in supporting better performance in linguistic tasks as compared to bilinguals. On the other hand, the fact that regions important in executive control mechanisms (i.e. IFG) are involved in these subnetworks (García-Pentón et al., 2014; Luk et al., 2011) is in line with Abutalebi and Green's model. Be that as it may, further complex brain network studies are needed to understand how over-developed structural and functional subnetworks and the entire functioning network correlate with executive control and behavioural linguistic tasks, respectively, in order to clear up the controversies surrounding the bilingual behavioural data. Particularly, it seems crucial to investigate which kinds of bilinguals and which conditions give rise to this cognitive advantage since there is very strong evidence showing no enhanced executive control in some bilinguals (Antón et al., 2014; Duñabeitia et al., 2014; Gathercole et al., 2014).

Assessing these findings in the light of Abutalebi and Green's Adaptive Control Hypothesis (Abutalebi & Green, 2007; Green & Abutalebi, 2013), there is just one region predicted by the model that consistently shows up across studies as a structural difference due to bilingualism: the left/right IFG (Hosoda et al, 2013; Klein et al., 2014; Grogan et al, 2012; Luk et al., 2011; García-Pentón et al., 2014). Some of the studies used alternatives to the traditional methods of VBM and ROI-based analysis and analyzed the whole brain to reveal effects of bilingualism in the IFG and in the connections between there and other regions (Luk et al., 2011; Klein et al., 2014; García-Pentón et al., 2014). Furthermore, others studies support these results showing that this region is also sensitive to L2 AoA and proficiency (Grogan et al., 2012; Hosoda et al., 2013; Klein et al., 2014). Additionally, several WM structural studies also confirmed differences between bilinguals and monolinguals in the tracts connecting IFG with many other regions in the frontal lobe (including the contralateral side) and the temporal, parietal and occipital regions in the back of the brain, specifically the CC (Luk et al., 2011; Pliatsikas et al., 2015; Gold et al., 2013; Mohades et al., 2012) and the IFOF (Luk et al., 2011; Pliatsikas et al., 2015; Mohades et al., 2012; Gold et al., 2013; Cummine & Boliek, 2013). Nevertheless, although these results identified the same regions, they are contradictory because some show increases while others show decreases in the WM.

In the same vein, some other regions predicted by the model also appeared. This is the case for the IPL, which was initially demonstrated by Mechelli et al. (2004) and then replicated by Grogan et al. (2012) and Abutalebi, Canini et al. (2015) but using a different methodology (namely, ROI-based rather than a whole-brain approach). And also for the ACC (Abutalebi, Guidi et al., 2015) but replication to support this finding is lacking.

Additionally, some regions emerged that are not predicted by the model: the aITG (Abutalebi et al., 2014) and cerebellum (Pliatsikas et al., 2014). Although the Adaptive Control Hypothesis could possibly account for the cerebellum in the context of dense code-switching, the sample in the Pliatsikas et al. (2014) study did not come from this sort of environment: the bilinguals in the study were in a relatively monolingual immersion setting, and more critically, their mean L2 AoA was 7.7 years old.

Although current neuronal models of bilingualism, such as Abutalebi and Green's Adaptive Control Hypothesis, are logical and consistent with the functional data, the current structural evidence does not provide complete support for the models' postulates. While the structural results offer limited support for some aspects of the Adaptive Control Hypothesis, taken as a whole they suggest that the model is incomplete and requires adjusting for those regions that cannot be accounted for or are altogether unexpected under its present formulation.

More critically, the current experimental evidence for plastic changes in the brain due to bilingual experience is relatively weak. Neuroimaging studies in this area are still very small in number and far from being consistent enough. With the evidence currently available, it is possible to be confident about consistent and reproducible structural changes related to bilingualism in only a few regions, such as IFG and connections with other areas. The remaining findings provide an unclear picture that makes it difficult to arrive at generalizations or to confirm or refute current models. Therefore, the debate over bilingual advantage does not seem to become clearer with current neuroimaging data. To a great extent this, much of this lack of conclusive evidence is due methodological differences among the studies and these inconsistencies will be identified in the next section, and

various solutions will be proposed to improve the situation. Against this backdrop, the Adaptive Control Hypothesis is a good candidate for a working model that further structural, functional and behavioural evidence would allow us to confirm and/or fine-tune. For that to happen, we need new studies with larger and better-documented samples – and preferably longitudinal designs – in order to accumulate more stable data.

Methodological concerns and recommendations

Various methodological issues have already been touched upon in the previous sections. One of the major concerns is about the different approaches used for the preprocessing and analysis of data, which can give rise to different results (Ashburner & Friston, 2011) and thus could contribute toward explaining the inconsistencies in the field of bilingualism. Tables 1-2 summarize the preprocessing and analysis of the studies included in this review. Notice that the greatest variation exists in GM studies, particularly in segmentation and registration procedures. Some of the studies used the unified segmentation approach (Ashburner & Friston, 2005), others used the improved unified segmentation approach implemented in the New Segment toolbox for SPM8 or even older segmentation algorithms (Ashburner & Friston, 2000), and yet others used the VBM5 protocol that does not use prior tissue information for the segmentation step (Good et al., 2001). There is also one study using the segmentation approach implemented in the FSL software (Zhang, Brady, & Smith, 2001), which relies on different algorithms. The registration step also depends on the software used for the processing of the images (see Table 1): old versions of SPM use a low spatial resolution method for the non-linear registration; FSL uses a medium spatial resolution method; and SPM8 uses a high spatial resolution registration method. Each of these methods can produce different results (Radua,

Canales-Rodriguez, Pomarol-Clotet, & Salvador, 2014). The size of the filter used to smooth the images also affects results and can be an important source of variability (Jones, Symms, Cercignani, & Howard, 2005; Salmond et al., 2002). In this case, no variability was observed across studies, with the single exception of the Abutalebi et al. (2013) study, which used an isotropic Gaussian kernel of 4mm of FWHM, compared to the other studies, which used 8mm (FWHM) or sigma of 3mm (approximately equivalent to 8mm FWHM). Additionally, although almost all of these studies used volume (modulated images) as the GM measure, a few studies used density (unmodulated images) as the GM measure (see Appendix 1). These different choices can give rise to different results and require different interpretations (Radua et al., 2014). Also, although some studies corrected images for brain size using the total intracranial volume (TIV), WM+GM raw volumes or age, others followed different statistical procedures or did not correct for brain size at all. Additionally, some studies used their own group template for the registration to the standard space, using DARTEL (Ashburner, 2007), FSL or other methods, while the rest of the studies used standard templates for registration. The former usually improves the registration in the group. These methodological choices need to be considered when interpreting the results and accounting for their variability.

In contrast, there is no such heterogeneity of methods in WM studies because – with the exception of only two studies that used a TOI approach, which suffers from the same shortcomings as the ROI approach (Furutani, Harada, Minato, Morita, & Nishitani, 2005; Kanaan et al., 2006; Snook, Plewes, & Beaulieu, 2007; Tapp et al., 2006) – the remaining studies follow a standardized method implemented in FSL to perform the TBSS approach on diffusion-derived measures. However, it is important take into account that the use of

data acquired from different scanners (1.5T or 3.0T) and using different parameters for the acquisition of the images (see tables 1 and 2 for details of the different studies reviewed here) can produce important differences in the quality of the images across studies and also influence variability in the results.

Finally, the most important difference between studies is the approach used for the statistical analysis of the data: ROI vs. VBM analysis, each of which answers different questions. The former looks for differences between groups at ROI-level and the latter looks for differences across the whole brain (i.e. voxel-level, peak-level, cluster-level inferences). Advantages and disadvantages of both procedures have been mentioned throughout this review, and there has been much debate about the use of voxel-based or ROI-based approaches (Good et al., 2001). Various studies have compared both methods (Furutani et al., 2005; Giuliani, Calhoun, Pearlson, Francis, & Buchanan, 2005; Kanaan et al., 2006; Kubicki et al., 2002; Snook et al., 2007; Suzuki et al., 2005; Tapp et al., 2006; Testa et al., 2004), finding, on the whole, similar results for both methods but some advantages for voxel-based over ROI-based analysis. Although VBM can overlook small differences (Saxe, Moran, Scholz, & Gabrieli, 2006), the ROI approach limits the chances of coming up with new, unexpected findings (Friston & Henson, 2006) making it difficult to expand and generalize the body of knowledge. Ultimately, the two techniques are complementary and cannot be used separately (Snook et al., 2007). This is important because even when there is a prior hypothesis about particular regions, it is useful to get information about the whole brain since regions are unlikely to be working in isolation, making it crucial to perform a whole brain analysis in the first place. Additionally, even though the ROI-based analysis increases the sensitivity of the test with respect to the

whole-brain analysis thanks to a reduction in the amount of testing and consequently in the problem of multiple comparisons, ROIs face other problems due to the effect of averaging discussed in section 2.1.1.2.

The VBM approach also has its weaknesses. The fact that different ways of performing VBM analysis can lead to disparate results is a huge problem for the integration of different studies. Different preprocessing steps applied to the images, such as the choice of segmentation or registration algorithms, or even the decision to modulate images (or not) after registration, can also lead to different results (Ridgway et al., 2008). However, ROI analysis holds no advantage over VBM in this respect because the definition of the regions by hand also introduces errors into the process. Usually, good segmentations are produced by well-trained and highly experienced research staff, and the difficulty lies in finding individuals with this kind of expertise. Furthermore, almost all the ROI studies reviewed here suffer from the same problem of image preprocessing as the VBM studies because they performed automatic extraction of the ROIs or they also normalize the images before the delineation of the ROIs by hand.

As far as data analysis is concerned, in order to make studies replicable and more generalizable, certain standards are needed (see Borgwardt, Radua, Mechelli, & Fusar-Poli, 2012; Ridgway et al., 2008, for comprehensive guidelines on good practice when reporting VBM studies). Here we wish to draw attention to three specific issues and to advocate a type of meta-analysis that would be particularly useful. Firstly, the various techniques of correcting for multiple comparisons require special attention and clarification: within the different methods of controlling the FWE rate or the FDR (described in Appendix 2), various options exist which impact on the interpretation of the results, and full details of the

correction process should be provided. Secondly, although reporting uncorrected results should be avoided, doing so requires providing even more information to make the results meaningful to interpretation (see Ridgway et al., 2008 for details of these recommendations). Thirdly, even if there is some justification for performing ROI analysis or SVC, a prior exploratory whole-brain analysis is needed to complement the ROI approach and even negative results must be reported (Borgwardt et al., 2012). This is crucial to get the full picture before reaching any conclusions.

With a view to bringing together evidence from different studies, voxel-based meta-analyses are the best quantitative tool to identify where differences in the brain really are, especially when the sample size of individual studies is a limitation (Borgwardt et al., 2012). This technique is even better than a standard qualitative review, because it makes it possible to obtain new p-values from many VBM studies. The problem here is that the small number of studies makes it impossible to perform this kind of analysis. Thus, more neuroimaging studies that include behavioural measures are needed. However, in the meantime, a database with the full statistical maps (not just the reported selections) of all the studies published so far could help to perform a meta-analysis that would clear up many issues and allow progress in the field.

In the same way that a correct and thoughtful methodological approach to data acquisition and analysis is important, so is the need for an adequate characterization of the bilingual sample being tested. Unfortunately, almost all the studies described in the current review have small samples and some of them provide minimal information regarding the type of bilinguals tested, thus making it difficult to draw conclusions that could generalize to other bilingual samples. (This may be another source of some of the abovementioned

discrepancies among existing studies.) The concept of bilingualism is broad enough to cover a finite but wide range of language combinations, and it has been established that languages of different typologies shape the human brain and its functions in different ways (Carreiras, Duñabeitia, Vergara, de la Cruz-Pavia, & Laka, 2010; Zhu, Nie, Chang, Gao, & Niu, 2014). Hence, it should be no surprise that different language combinations (pairs) give rise to differing development of the neural substrates that support language use and control. Similarly, even if the results obtained from studies exploring the influence of L2 proficiency and L2 AoA partially converge, there is convincing evidence that these two factors independently contribute to language processing in bilinguals' comprehension and production behaviour (Dimitropoulou, Duñabeitia, & Carreiras, 2011; Dowens, Vergara, Barber, & Carreiras, 2010; Duñabeitia, Dimitropoulou, Uribe-Etxebarria, Laka, & Carreiras, 2010; Duñabeitia, Perea, & Carreiras, 2010; Perea, Duñabeitia, & Carreiras, 2008). Thus, a thorough description of the knowledge and use of each of the languages is essential for a precise characterization of the samples being tested to make possible the replication and discussion of findings in the context of the specific linguistic background of the participants (see Tables 1 and 2 for an illustration of the variability between studies). Finally, but equally important, a precise definition of the manner in which the second language has been acquired and of the context in which each language is being used is critical, given numerous demonstrations of the differential effects derived from naturalistic vs. classroom-based learning (Muñoz, 2008; Pliatsikas & Marinis, 2013; see Stein, Winkler, Kaiser, & Dierks, 2014, for an overview), as well as of dominance-switch effects (Basnight-Brown & Altarriba, 2007), and of the role of immersion in language processing, and therefore, in the neural assemblies supporting bilingualism (Baus, Costa, & Carreiras, 2013). Since bilingualism is in and of itself a heterogeneous phenomenon, a wide range of

studies taking in this variability could reduce this methodological problem and provide better answers. Unfortunately, to date there are insufficient studies covering and replicating this variability.

In sum, more studies with higher numbers of participants and a better description of the samples and methodology are needed to accumulate an important body of evidence to illuminate whether and how bilingualism modulates brain structure and function and to obtain more stable results across studies. In addition, there is a need for more research combining behavioural and brain measures to understand among other things (1) how potential brain changes in specific areas/circuits (which should be replicated by several studies) are related to cognitive processes and behaviour, (2) how and whether these brain changes are modulated by AoA, proficiency and language combinations, and (3) whether bilingual advantages at the behavioural level are accompanied by observable changes in the brain with the aim of understanding when these bilingual advantages do or do not appear and why.

Closing remarks

So far, the picture is still too blurry and definitely less clear than expected, with very few data points and not enough consistent findings across studies. As a note of caution, it is also important to keep in mind that brain evidence is much wanted, but does not provide the definitive answer: neuroimaging data are no more than another important piece of information helping to solve the puzzle. Simply changing the nature of the data, that is, moving from behavioural data to structural brain data, will not solve the debate of bilingual advantage. There is a need for further studies that combine behavioural and brain measures in a systematic and transparent manner, providing clear and complete descriptions

of sample characteristics and data processing and analysis. Developing models that respond to the data and take in the multiple factors that make up the bilingual panorama will make it possible to test hypotheses and tease apart the roles those factors play in shaping the bilingual brain. Only then will we be able to shed light on what, if any, brain circuits change with learning two languages in different situations and how the brain and the cognitive system negotiate the processing of two languages.

Acknowledgments: The authors are very grateful to Erick J. Canales and Yasser Iturria-Medina for their comments and revision of this manuscript. This article was partially supported by grants CONSOLIDER-INGENIO2010 CSD2008-00048, PSI2012-32123 and PSI2012-31448 from the Spanish Ministry of Science and Innovation, by ERC-2011-ADG-295362 grant from the European Research Council and by the AThEME project funded by the European Union Seventh Framework Programme (grant number 613465). The authors declare no competing interests.

6. References

Abutalebi, J., & Green, D. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics, 20*(3), 242-275

Abutalebi, J., Canini, M., Della Rosa, P. A., Green, D. W., & Weekes, B. S. (2015). The neuroprotective effects of bilingualism upon the inferior parietal lobule: a structural neuroimaging study in aging Chinese bilinguals. *Journal of Neurolinguistics*, *33*, 3-13. doi: 10.1016/j.jneuroling.2014.09.008

Abutalebi, J., Canini, M., Della Rosa, P. A., Sheung, L. P., Green, D. W., & Weekes, B. S. (2014). Bilingualism protects anterior temporal lobe integrity in aging. *Neurobiology of Aging, 35*(9), 2126-2133. doi: 10.1016/j.neurobiolaging.2014.03.010

Abutalebi, J., Della Rosa, P. A., Gonzaga, A. K., Keim, R., Costa, A., & Perani, D. (2013). The role of the left putamen in multilingual language production. *Brain and Language, 125*(3), 307-315. doi: 10.1016/j.bandl.2012.03.009

Abutalebi, J., Guidi, L., Borsa, V., Canini, M., Della Rosa, P. A., Parris, B. A., & Weekes, B. S. (2015). Bilingualism provides a neural reserve for aging populations. *Neuropsychologia*, *69*, 201-210. doi: 10.1016/j.neuropsychologia.2015.01.040

Antón, E., Duñabeitia, J. A., Estévez, A., Hernández, J. A., Castillo, A., Fuentes, L. J., . . . Carreiras, M. (2014). Is there a bilingual advantage in the ANT task? Evidence from children. *Frontiers in Psychology, 5*, 398. doi: 10.3389/fpsyg.2014.00398

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage, 38*(1), 95-113. doi: 10.1016/j.neuroimage.2007.07.007

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry--the methods. *Neuroimage, 11*(6 Pt 1), 805-821. doi: 10.1006/nimg.2000.0582

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage, 26*(3), 839-851. doi: 10.1016/j.neuroimage.2005.02.018

Ashburner, J., & Friston, K. J. (2011). Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. *Neuroimage, 55*(3), 954-967. doi: 10.1016/j.neuroimage.2010.12.049

Basnight-Brown, D. M., & Altarriba, J. (2007). Differences in semantic and translation priming across languages: The role of language direction and language dominance. *Memory & Cognition, 35*(5), 953-965

Basser, P. J., Mattiello, J., & LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical Journal, 66*(1), 259-267. doi: 10.1016/s0006-3495(94)80775-1

Baus, C., Costa, A., & Carreiras, M. (2013). On the effects of second language immersion on first language production. *Acta Psychologica, 142*(3), 402-409

Bialystok, E., & Barac, R. (2012). Emerging bilingualism: dissociating advantages for metalinguistic awareness and executive control. *Cognition, 122*(1), 67-73. doi: 10.1016/j.cognition.2011.08.003

Bialystok, E., Craik, F. I., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in Cognitive Sciences, 16*(4), 240-250. doi: 10.1016/j.tics.2012.03.001

Bialystok, E., Kroll, J.F., Green, D.W., MacWhinney, B., Craik, F.I. (2015). Publication Bias and the Validity of Evidence: What's the Connection? *Psychol Sci.* 26(6), 944-6. doi: 10.1177/0956797615573759

Borgwardt, S., Radua, J., Mechelli, A., & Fusar-Poli, P. (2012). Why are psychiatric imaging methods clinically unreliable? Conclusions and practical guidelines for authors, editors and reviewers. *Behavioral and Brain Functions, 8*, 46. doi: 10.1186/1744-9081-8-46

Carreiras, M., Duñabeitia, J. A., Vergara, M., de la Cruz-Pavia, I., & Laka, I. (2010). Subject relative clauses are not universally easier to process: Evidence from Basque. *Cognition, 115*(1), 79-92. doi: 10.1016/j.cognition.2009.11.012

Carreiras, M., Seghier, M. L., Baquero, S., Estévez, A., Lozano, A., Devlin, J. T., & Price, C. J. (2009). An anatomical signature for literacy. *Nature, 461*(7266), 983-986. doi: 10.1038/nature08461

Chung, M. K., Dalton, K. M., Shen, L., Evans, A. C., & Davidson, R. J. (2007). Weighted fourier series representation and its application to quantifying the amount of gray matter. *IEEE Transactions on Medical Imaging, 26*(4), 566-581. doi: 10.1109/tmi.2007.892519

Clare, L., Whitaker, C.J., Craik, F.I., Bialystok, E., Martyr, A., Martin-Forbes, P.A., . . . Hindle, J.V. (2014). Bilingualism, executive control, and age at diagnosis among people with early-stage Alzheimer's disease in Wales. *J Neuropsychol*. doi: 10.1111/jnp.12061

Costa, A., & Sebastián-Gallés, N. (2014). How does the bilingual experience sculpt the brain?. *Nature Reviews Neuroscience, 15, 336–345*. doi:10.1038/nrn3709

Costa, A., Hernández, M., & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: evidence from the ANT task. *Cognition, 106*(1), 59-86. doi: 10.1016/j.cognition.2006.12.013

Costa, A., Hernández, M., Costa-Faidella, J., & Sebastián-Gallés, N. (2009). On the bilingual advantage in conflict processing: now you see it, now you don't. *Cognition, 113*(2), 135-149. doi: 10.1016/j.cognition.2009.08.001

Crystal, D. (1997). *English as a global language*: Cambridge University Press.

Cummine, J., & Boliek, C. A. (2013). Understanding white matter integrity stability for bilinguals on language status and reading performance. *Brain Structure & Function, 218*(2), 595-601. doi: 10.1007/s00429-012-0466-6

De Baene, W., Duyck, W., Brass, M., & Carreiras, M. (2015). Brain circuit for cognitive control is shared by task and language switching. *Journal of Cognitive Neuroscience, doi:10.1162/jocn_a_00817*

de Bruin, A., Treccani, B., & Della Sala, S. (2015a). Cognitive advantage in bilingualism an example of publication bias? *Psychological Science*, *26*(1), 99-107

de Bruin, A., Treccani, B, Della Sala, S. (2015b). The connection is in the data: we should consider them all. *Psychol Sci.* 26(6), 947-9. doi: 10.1177/0956797615583443

Dimitropoulou, M., Duñabeitia, J. A., & Carreiras, M. (2011). Two words, one meaning: evidence of automatic co-activation of translation equivalents. *Frontiers in Psychology, 2*, 188. doi: 10.3389/fpsyg.2011.00188

Dowens, M. G., Vergara, M., Barber, H. A., & Carreiras, M. (2010). Morphosyntactic processing in late second-language learners. *Journal of Cognitive Neuroscience, 22*(8), 1870-1887

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., & May, A. (2004). Neuroplasticity: changes in grey matter induced by training. *Nature, 427*(6972), 311-312. doi: 10.1038/427311a

Duñabeitia, J. A., Dimitropoulou, M., Uribe-Etxebarria, O., Laka, I., & Carreiras, M. (2010). Electrophysiological correlates of the masked translation priming effect with

highly proficient simultaneous bilinguals. *Brain Research, 1359*, 142-154. doi: 10.1016/j.brainres.2010.08.066

Duñabeitia, J. A., Hernández, J. A., Antón, E., Macizo, P., Estévez, A., Fuentes, L. J., & Carreiras, M. (2014). The inhibitory advantage in bilingual children revisited: Myth or reality? *Experimental Psychology, 61*(3), 234-251

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2010). Masked translation priming effects with highly proficient simultaneous bilinguals. *Experimental Psychology, 57*(2), 98-107. doi: 10.1027/1618-3169/a000013

Duñabeitia, J.A., & Carreiras, M. (in press). The bilingual advantage: acta est fabula? *Cortex*

Duñabeitia, J.A., Fernández, Y., & Carreiras, M. (submitted). Does bilingualism shape inhibitory control in the elderly?

Edwards, J. (2004). Foundations of Bilingualism. In T. K. B. W. Ritchie (Ed.), *The Handbook of Bilingualism* (pp. 7–31): Oxford, U.K.: Blackwell

Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America, 97*(20), 11050-11055. doi: 10.1073/pnas.200033797

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, 137, 201–225

Friston, K. J., & Henson, R. N. (2006). Commentary on: divide and conquer; a defence of functional localisers. *Neuroimage, 30*(4), 1097-1099

Furutani, K., Harada, M., Minato, M., Morita, N., & Nishitani, H. (2005). Regional changes of fractional anisotropy with normal aging using statistical parametric mapping (SPM). *The Journal of Medical Investigation: JMI, 52*(3-4), 186-190

García-Pentón, L., Pérez Fernández, A., Iturria-Medina, Y., Gillon-Dowens, M., & Carreiras, M. (2014). Anatomical connectivity changes in the bilingual brain. *Neuroimage, 84*, 495-504. doi: 10.1016/j.neuroimage.2013.08.064

Gaser, C., & Schlaug, G. (2003). Brain structures differ between musicians and non-musicians. *The Journal of Neuroscience, 23*(27), 9240-9245

Gathercole, V. C., Thomas, E. M., Kennedy, I., Prys, C., Young, N., Vinas Guasch, N., . . . Jones, L. (2014). Does language dominance affect cognitive performance in bilinguals? Lifespan evidence from preschoolers through older adults on card sorting, Simon, and metalinguistic tasks. *Frontiers in Psychology, 5*, 11. doi: 10.3389/fpsyg.2014.00011

Giuliani, N. R., Calhoun, V. D., Pearlson, G. D., Francis, A., & Buchanan, R. W. (2005). Voxel-based morphometry versus region of interest: a comparison of two

methods for analyzing gray matter differences in schizophrenia. *Schizophrenia Research, 74*(2-3), 135-147. doi: 10.1016/j.schres.2004.08.019

Gold, B. T., Johnson, N. F., & Powell, D. K. (2013). Lifelong bilingualism contributes to cognitive reserve against white matter integrity declines in aging. *Neuropsychologia, 51*(13), 2841-2846. doi: 10.1016/j.neuropsychologia.2013.09.037

Gollan, T. H., Sandoval, T., & Salmon, D. P. (2011). Cross-language intrusion errors in aging bilinguals reveal the link between executive control and language selection. *Psychological Science, 22*(9), 1155-1164. doi: 10.1177/0956797611417002

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Friston, K. J., & Frackowiak, R. S. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage, 14*(1 Pt 1), 21-36. doi: 10.1006/nimg.2001.0786

Green, D.W., & Abutalebi, J. (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology, 25(5)*, 515-530. doi: 10.1080/20445911.2013.796377

Grogan, A., Parker Jones, O., Ali, N., Crinion, J., Orabona, S., Mechias, M. L., . . . Price, C. J. (2012). Structural correlates for lexical efficiency and number of languages in non-native speakers of English. *Neuropsychologia, 50*(7), 1347-1352. doi: 10.1016/j.neuropsychologia.2012.02.019

Grosjean, F. (1998). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition, 1*(02), 131-149

Grosjean, F. (2010). *Bilingual: Life and reality.* Harvard University Press.

Guye, M., Bartolomei, F. and Ranjeva, J. P. (2008). Imaging structural and functional connectivity: towards a unified definition of human brain organization? *Curr Opin Neurol.* 21(4), 393-403

Hernandez, A. E. (2009). Language switching in the bilingual brain: What's next?. *Brain and Language*, *109*(2), 133-140

Hernandez, A. E. (2013). *The bilingual brain*: Oxford University Press.

Hernandez, A. E., & Li, P. (2007). Age of acquisition: its neural and computational mechanisms. *Psychological Bulletin*, *133*(4), 638-650. doi: 10.1037/0033-2909.133.4.638

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, *9*(5), 220-225. doi:10.1016/j.tics.2005.03.003

Hosoda, C., Tanaka, K., Nariai, T., Honda, M., & Hanakawa, T. (2013). Dynamic neural network reorganization associated with second language vocabulary acquisition: a multimodal imaging study. *The Journal of Neuroscience, 33*(34), 13663-13672. doi: 10.1523/jneurosci.0410-13.2013

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). FSL. *Neuroimage, 62*(2), 782-790. doi: 10.1016/j.neuroimage.2011.09.015

Jones, D. K., Knosche, T. R., & Turner, R. (2013). White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI. *Neuroimage, 73*, 239-254. doi: 10.1016/j.neuroimage.2012.06.081

Jones, D. K., Symms, M. R., Cercignani, M., & Howard, R. J. (2005). The effect of filter size on VBM analyses of DT-MRI data. *Neuroimage, 26*(2), 546-554. doi: 10.1016/j.neuroimage.2005.02.013

Kanaan, R. A., Shergill, S. S., Barker, G. J., Catani, M., Ng, V. W., Howard, R., . . . Jones, D. K. (2006). Tract-specific anisotropy measurements in diffusion tensor imaging. *Psychiatry Research, 146*(1), 73-82. doi: 10.1016/j.pscychresns.2005.11.002

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience, 12*(4), 231-242. doi: 10.1038/nrn3000

Kennedy, D., & Norman, C. (2005). What Don't We Know? *Science, 309*(5731), 75. doi: 10.1126/science.309.5731.75

Klein, D., Mok, K., Chen, J. K., & Watkins, K. E. (2014). Age of language learning shapes brain structure: a cortical thickness study of bilingual and monolingual individuals. *Brain and Language, 131*, 20-24. doi: 10.1016/j.bandl.2013.05.014

Kroll, J. F., & Bialystok, E. (2013). Understanding the consequences of bilingualism for language processing and cognition. *Journal of Cognitive Psychology*, *25*(5), 497-514. doi: 10.1080/20445911.2013.799170

Kubicki, M., Shenton, M. E., Salisbury, D. F., Hirayasu, Y., Kasai, K., Kikinis, R., . . . McCarley, R. W. (2002). Voxel-based morphometric analysis of gray matter in first episode schizophrenia. *Neuroimage, 17*(4), 1711-1719.

Lee, H., Devlin, J. T., Shakeshaft, C., Stewart, L. H., Brennan, A., Glensman, J., . . . Price, C. J. (2007). Anatomical traces of vocabulary acquisition in the adolescent brain. *The Journal of Neuroscience, 27*(5), 1184-1189. doi: 10.1523/jneurosci.4442-06.2007

Li, P., Legault, J., & Litcofsky, K. A. (2014). Neuroplasticity as a function of second language learning: anatomical changes in the human brain. *Cortex, 58*, 301-324. doi: 10.1016/j.cortex.2014.05.001

Luk, G., Bialystok, E., Craik, F. I., & Grady, C. L. (2011). Lifelong bilingualism maintains white matter integrity in older adults. *The Journal of Neuroscience, 31*(46), 16808-16813. doi: 10.1523/jneurosci.4563-11.2011

Luk, G., Green, D. W., Abutalebi, J., & Grady, C. (2012). Cognitive control for language switching in bilinguals: A quantitative meta-analysis of functional neuroimaging studies. *Language and Cognitive Processes*, *27*(10), 1479-1488. doi:10.1080/01690965.2011.613209

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, *97*(8), 4398-4403.

Maguire, E. A.,Woollett, K. and Spiers, H. J. (2006). London taxi drivers and bus drivers: a structural MRI and neuropsychological analysis. *Hippocampus*, 16(12), 1091-1101.

Mårtensson, J., Eriksson, J., Bodammer, N. C., Lindgren, M., Johansson, M., Nyberg, L., & Lovden, M. (2012). Growth of language-related brain areas after foreign language learning. *Neuroimage, 63*(1), 240-244. doi: 10.1016/j.neuroimage.2012.06.043

Martin, C. D., Costa, A., Dering, B., Hoshino, N., Wu, Y. J., & Thierry, G. (2012). Effects of speed of word processing on semantic access: the case of bilingualism. *Brain and Language, 120*(1), 61-65. doi: 10.1016/j.bandl.2011.10.003

Mechelli, A., Crinion, J. T., Noppeney, U., O'Doherty, J., Ashburner, J., Frackowiak, R. S., & Price, C. J. (2004). Neurolinguistics: structural plasticity in the bilingual brain. *Nature, 431*(7010), 757. doi: 10.1038/431757a.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49-100. doi:10.1006/cogp.1999.0734.

Miyake, A. & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychology*, 21(1), 8–14.

Mohades, S. G., Struys, E., Van Schuerbeek, P., Mondt, K., Van De Craen, P., & Luypaert, R. (2012). DTI reveals structural differences in white matter tracts between bilingual and monolingual children. *Brain Research, 1435*, 72-80. doi: 10.1016/j.brainres.2011.12.005

Mohades, S. G., Van Schuerbeek, P., Rosseel, Y., Van De Craen, P., Luypaert, R., & Baeken, C. (2015). White-Matter development is different in bilingual and monolingual children: a Longitudinal DTI study. *PloS one*, *10*(2). doi: 10.1371/journal.pone.0117968

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics, 29*(4), 578-596.

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage, 62*(2), 811-815. doi: 10.1016/j.neuroimage.2012.04.014

Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research, 12*(5), 419-446.

Olsen, R. K., Pangelinan, M. M., Bogulski, C., Chakravarty, M. M., Luk, G., Grady, C., & Bialystok, E. (2015). The effect of lifelong bilingualism on regional grey and white matter volume. *Brain Research*. Available online 25 February 2015. doi: 10.1016/j.brainres.2015.02.034.

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology, 66*(2), 232-258.

Paap, K. R., & Sawi, O. (2014). Bilingual advantages in executive functioning: problems in convergent validity, discriminant validity, and the identification of the theoretical constructs. *Frontiers in Psychology, 5*, 962.

Paap, K. R., Johnson, H. A., & Sawi, O. (2014). Are bilingual advantages dependent upon specific tasks or specific bilingual experiences? *Journal of Cognitive Psychology, ahead-of-print*, 1-25.

Paap, K.R., Johnson, H.A., & Sawi, O. (2015). Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. *Cortex*. doi: 10.1016/j.cortex.2015.04.014

Perea, M., Duñabeitia, J. A., & Carreiras, M. (2008). Masked associative/semantic priming effects across languages with highly proficient bilinguals. *Journal of Memory and Language, 58*(4), 916-930.

Pliatsikas, C., & Marinis, T. (2013). Processing of regular and irregular past tense morphology in highly proficient second language learners of English: a self-paced reading study. *Applied Psycholinguistics, 34*(05), 943-970.

Pliatsikas, C., Johnstone, T., & Marinis, T. (2014). Grey matter volume in the cerebellum is related to the processing of grammatical rules in a second language: a structural voxel-based morphometry study. *Cerebellum, 13*(1), 55-63. doi: 10.1007/s12311-013-0515-6

Pliatsikas, C., Moschopoulou, E., & Saddy, J. D. (2015). The effects of bilingualism on the white matter structure of the brain. *Proceedings of the National Academy of Sciences*, 112 (5), 1334–1337. doi:10.1073/pnas.1414183112

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences, 1191*, 62-88. doi: 10.1111/j.1749-6632.2010.05444.x

Radua, J., Canales-Rodriguez, E. J., Pomarol-Clotet, E., & Salvador, R. (2014). Validity of modulation and optimal settings for advanced voxel-based morphometry. *Neuroimage, 86*, 81-90. doi: 10.1016/j.neuroimage.2013.07.084

Ressel, V., Pallier, C., Ventura-Campos, N., Díaz, B., Roessler, A., Ávila, C., & Sebastián-Gallés, N. (2012). An effect of bilingualism on the auditory cortex. *The Journal of Neuroscience, 32*(47), 16597-16601.

Ridgway, G. R., Henley, S. M., Rohrer, J. D., Scahill, R. I., Warren, J. D., & Fox, N. C. (2008). Ten simple rules for reporting voxel-based morphometry studies. *Neuroimage, 40*(4), 1429-1435. doi: 10.1016/j.neuroimage.2008.01.003

Salat, D. H., Tuch, D. S., Greve, D. N., van der Kouwe, A. J., Hevelone, N. D., Zaleta, A. K., . . . Dale, A. M. (2005). Age-related alterations in white matter microstructure measured by diffusion tensor imaging. *Neurobiology of Aging, 26*(8), 1215-1227. doi: 10.1016/j.neurobiolaging.2004.09.017

Salmond, C. H., Ashburner, J., Vargha-Khadem, F., Connelly, A., Gadian, D. G., & Friston, K. J. (2002). Distributional assumptions in voxel-based morphometry. *Neuroimage, 17*(2), 1027-1030.

Saxe, R., Moran, J. M., Scholz, J., & Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience, 1*(3), 229-234. doi: 10.1093/scan/nsl034

Schlegel, A. A., Rudelson, J. J., & Tse, P. U. (2012). White matter structure changes as adults learn a second language. *Journal of Cognitive Neuroscience, 24*(8), 1664-1670. doi: 10.1162/jocn_a_00240

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage, 44*(1), 83-98. doi: 10.1016/j.neuroimage.2008.03.061

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., . . . Behrens, T. E. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage, 31*(4), 1487-1505. doi: 10.1016/j.neuroimage.2006.02.024

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., . . . Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage, 23 Suppl 1*, S208-S219. doi: 10.1016/j.neuroimage.2004.07.051

Snook, L., Plewes, C., & Beaulieu, C. (2007). Voxel based versus region of interest analysis in diffusion tensor imaging of neurodevelopment. *Neuroimage, 34*(1), 243-252. doi: 10.1016/j.neuroimage.2006.07.021

Stein, M., Federspiel, A., Koenig, T., Wirth, M., Strik, W., Wiest, R., . . . Dierks, T. (2012). Structural plasticity in the language system related to increased second language proficiency. *Cortex, 48*(4), 458-465.

Stein, M., Winkler, C., Kaiser, A. C., & Dierks, T. (2014). Structural brain changes related to bilingualism: Does immersion make a difference? *Frontiers in Psychology, 5*, 1116. doi: 10.3389/fpsyg.2014.01116

Suzuki, M., Hagino, H., Nohara, S., Zhou, S. Y., Kawasaki, Y., Takahashi, T., . . . Kurachi, M. (2005). Male-specific volume expansion of the human hippocampus during adolescence. *Cerebral Cortex, 15*(2), 187-193. doi: 10.1093/cercor/bhh121

Tapp, P. D., Head, K., Head, E., Milgram, N. W., Muggenburg, B. A., & Su, M. Y. (2006). Application of an automated voxel-based morphometry technique to assess regional gray and white matter brain atrophy in a canine model of aging. *Neuroimage, 29*(1), 234-244. doi: 10.1016/j.neuroimage.2005.07.043

Testa, C., Laakso, M. P., Sabattoli, F., Rossi, R., Beltramello, A., Soininen, H., & Frisoni, G. B. (2004). A comparison between the accuracy of voxel-based morphometry and hippocampal volumetry in Alzheimer's disease. *Journal of Magnetic Resonance Imaging, 19*(3), 274-282. doi: 10.1002/jmri.20001

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., . . . Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage, 15*(1), 273-289. doi: 10.1006/nimg.2001.0978

Winkler, A. M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P. T., . . . Glahn, D. C. (2010). Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *Neuroimage, 53*(3), 1135-1146. doi: 10.1016/j.neuroimage.2009.12.028

Zalesky, A., Fornito, A., & Bullmore, E. T. (2010). Network-based statistic: identifying differences in brain networks. *Neuroimage, 53*(4), 1197-1207. doi: 10.1016/j.neuroimage.2010.06.041

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging, 20*(1), 45-57. doi: 10.1109/42.906424

Zhu, L., Nie, Y., Chang, C., Gao, J. H., & Niu, Z. (2014). Different patterns and development characteristics of processing written logographic characters and alphabetic words: an ALE meta-analysis. *Human Brain Mapping, 35*(6), 2607-2618. doi: 10.1002/hbm.22354

Zou, L., Ding, G., Abutalebi, J., Shu, H., & Peng, D. (2012). Structural plasticity of the left caudate in bimodal bilinguals. *Cortex, 48*(9), 1197-1206. doi: 10.1016/j.cortex.2011.05.022

Appendix 1. Techniques and measures used in structural brain analysis

Voxel-based morphometry (VBM) is a whole-brain technique that allows the investigation of local differences in the brain using statistical parametric mapping (SPM). VBM requires T1-weighted MRI images to be registered to a template and classified into three different brain tissue classes: grey matter (GM), white matter (WM) and cerebral

spinal fluid (CSF). The segmentation into different brain tissue classes can be done in one step by combining tissue classifications using image voxel intensities, bias field correction and the prior probability derived from registration to a set of a priori tissue probability maps (e.g. GM, WM, CSF). These steps are integrated in the same generative model, known as unified segmentation (Ashburner & Friston, 2005), which solved the circularity of previous registration-segmentation procedures (Ashburner & Friston, 2000). After spatial registration, images are typically scaled to compensate for any contraction during registration (known as modulation), thus conserving the total amount of GM/WM/CSF as in the original images. By this means, volumetric differences can be tested for, which means that not only mesoscopic (i.e. between microscopic and macroscopic) regional changes in the brain, such as cortical thinning, can be detected, but also macroscopic regional changes, such as cortical folding (Radua et al., 2014). If the images are not corrected (unmodulated) concentration or density differences alone can be tested for, making it possible to detect only mesoscopic differences (Radua et al., 2014). Subsequently, images must be spatially smoothed to permit the comparison of the volume/density images across brains for each individual voxel, using the general linear model (Ashburner & Friston, 2000). Because this model involves the use of univariate statistics at each voxel, many statistical tests are conducted. Therefore, the statistical significance of the inferences must be adjusted to correct for the problem of multiple comparisons (see Appendix 2).

Region of interest (ROI) approaches restrict the statistical analysis to a specific region or regions, which may be defined by manually drawing the limits on the individual native space or by automatic parcellation, which involves segmentation and registration

preprocesses and then individual atlas labelling using a standardized atlas to demarcate the different anatomical structures or regions. The desired measure (e.g. GM volume or density) is extracted from the images and averaged to obtain a global measure for each region under consideration.

Cortical thickness (CT) measurement involves registration to the standard space, tessellation of the GM and WM boundaries, automated topology correction and surface deformation following intensity gradients to optimally place the GM/WM and GM/CSF borders at the location where the greatest shift in intensity defines the transition between the different classes of tissue. Deformation procedures include surface inflation and registration to a spherical atlas. The method uses both intensity and continuity information from the entire three-dimensional T1-weighted MRI in the segmentation and deformation procedures to produce representations of the cortical thickness, calculated as the closest distance from the GM/WM boundary to the GM/CSF boundary at each vertex on the tessellated surface. The resulting maps are not restricted to the voxel resolution of the original data, and thus can detect sub-millimetre differences between groups. Before performing statistical analysis, the individual cortical thickness maps must be smoothed and finally a vertex-wise general linear model can be applied or a ROI approach can be used.

Diffusion tensor imaging (DTI) is estimated from the DW-MRI (Basser, Mattiello, & LeBihan, 1994), which measures the motion of the water molecules across the axon, providing information about the fibre orientation and organization. Then, scalar measures associated to each diffusion tensor are used to obtain invariant indices like the mean diffusivity (MD), which characterizes the overall water diffusion in each voxel of the brain

(for example, MD is higher in ventricles, lower in bones and tends to decrease with increases of myelination). Other scalar measures are the axial/radial diffusivity (AD/RD) that describe, respectively, water mobility along the axis of the main fibre orientation and water mobility perpendicular to this axis (Jones, Knosche, & Turner, 2013). Perhaps the most widely used tensor-derived measure is Fractional anisotropy (FA), which is calculated as the relationship between AD and RD measures and provides information about the degree of anisotropy of the water diffusion in the voxel. The anisotropy is higher (close to 1) inside the axon since the water is impeded from moving across the axon membrane (but can move more freely along the axon), and is lower (close to 0) in regions where the water can move freely in any direction, such as ventricles. Importantly, increasing axonal density, reducing axonal calibre or increasing the degree of myelination should all reduce RD and therefore elevate FA. Despite the extensive use of these measures in many fields of neuroscience, any differences in values should not always be associated with or interpreted in terms of WM tissue "integrity". Different fibre configurations and variations in these configurations can produce different modifications in these measures (Jones et al., 2013).

Tract-based spatial statistics (TBSS) uses an improved nonlinear registration procedure and a mean FA skeleton (which represents the centre of all common tracts) to project each subject's FA maps. This avoids data smoothing and increases the sensitivity of the voxel-wise cross-subject statistics. Another advantage of this technique is that it only examines areas where the fibres run parallel (i.e. the voxels inside the skeleton). These provide a better interpretation of the results, since in areas of crossing fibres the FA changes are more difficult to interpret in terms of WM volume or integrity.

Appendix 2. Corrections for multiple comparisons

The most commonly used method to correct for multiple comparisons is to control the **family wise error** rate (FWE) using random field theory and resampling-based approaches (Nichols, 2012; Nichols & Hayasaka, 2003) and this can be applied at the voxel-level or cluster-level of inference. In general, voxel-level FWE controlling procedures have good spatial specificity but poor sensitivity, and cluster-level FWE controlling procedures have better sensitivity but poor spatial specificity. More recently, **false discovery rate** (FDR) has been used to correct for the multiple comparisons problem at voxel-level. Which method is more appropriate and accurate depends on whether the data fulfil the assumptions of Gaussian distribution underlying each technique.

Small volume corrections (SVC) of the FWE limit the analyses to the scope of certain subvolume but without the averaging inherent to ROI approach. SVC makes it possible to correct for multiple comparisons based just on the number of voxels in the subvolume, which is a more liberal correction.

Appendix 3. Glossary. Abbreviations used in this review.

Note: Acronyms for neuroanatomical terms comply as far as possible to those used in NeuroNames available at http://braininfo.org (Bowden, Song, Kosheleva, & Dubach, 2012). Deviations occur in the capitalization of certain acronyms (e.g. 'CC' rather than 'cc' for corpus callosum) to aid text legibility.

ACC	anterior cingulate cortex
AD	axial diffusivity
aITG	anterior inferior temporal gyrus
AnG	angular gyrus
AoA	Age of acquisition

CC	corpus callosum
CSF	cerebral spinal fluid
CSL	Chinese Sign Language
CT	cortical thickness
DW-MRI	diffusion-weighted magnetic functional imaging
FA	fractional anisotropy
FDR	false discovery rate
FSL	FRMIB Software Library
FWE	family wise error rate
FWHM	full-width at half-maximum
GM	grey matter
IFG	inferior frontal gyrus
IFGOp	pars opercularis
IFGOr	pars orbitalis
IFGTr	pars triangularis
IFOF	inferior fronto-occipital fasciculus
ILF	inferior longitudinal fasciculus
IPL	inferior parietal lobule
L2	second language
MD	mean diffusivity
MFG	middle frontal gyrus
MRI	magnetic resonance imaging
NBS	network-based statistics
OFC	orbito-frontal cortex
RD	radial diffusivity

RFT	random field theory
ROI	region of interest
RS	resting state
SLF	superior longitudinal fasciculus
SMG	supramarginal gyrus
SPL	superior parietal lobule
SPM	statistical parameter mapping
STG	superior temporal gyrus
SVC	small volume correction
TBSS	tract-based spatial statistics
TFCE	threshold free cluster enhancement
TIV	total intracranial volume
TmP	temporal pole
TOI	tract of interest
VBM	voxel-based morphometry
WM	white matter

Figure Captions

Figure 1. Cross-sectional studies showing significant differences in GM volume/density in bilinguals as compared to monolinguals **a)** Results from VBM studies at whole-brain analysis. The colors represent the relative location in the brain of the results from different studies. ***Red***: Mechelli et al. (2004), showing significant GM density increase in left inferior parietal lobule [IPL]. ***Blue***: Pliatsikas et al. (2014), showing increased GM volume

in the cerebellum. ***Yellow***: Abutalebi et al. (2014), showing increased GM volume in left anterior inferior temporal gyrus [aITG]. ***Green***: Abutalebi, Guidi, et al. (2015), showing increased GM volume in right/left anterior cingulate cortex [ACC]. Abbreviations: R (right); L (Left); A (anterior); P (posterior). **b)** Results from VBM studies using small volume correction (SVC), these studies also showed higher GM volume/density in bilinguals as compared to monolinguals –with the exception of Abutalebi et al. (2013), which compared multilinguals vs. monolinguals, and Zou et al. (2012), which compared bimodal bilinguals vs. monolinguals. ***Red***: Ressel et al. (2012), showing increased GM volume in left Heschl gyrus. ***Blue***: Abutalebi et al. (2013), showing increased GM volume in left putamen. ***Purple***: Zou et al., (2012), showing increased volume in the left caudate. ***Green***: Grogan et al. (2012), showing increased GM volume in right IPL, using Mechelli et al.'s (2004) coordinates for the SVC. Abbreviations: R (right); L (Left); A (anterior); P (posterior).

Figure 2. Regions showing significant differences in FA values between bilinguals and monolinguals from TBSS studies. ***Green***: corpus callosum [CC] (Luk et al., 2011; Gold et al., 2013; Pliatsikas et al., 2015). ***Red:*** inferior frontal-occipital fasciculus [IFOF] (Luk et al., 2011; Cummine & Boliek, 2013; Gold et al., 2013; Pliatsikas et al., 2015). ***Blue***: superior longitudinal fasciculi [SLF] (Luk et al., 2011; Pliatsikas et al., 2015). Abbreviations: R (right); L (Left); A (anterior); P (posterior).

Table 1: Cross-sectional GM studies related to bilingualism (ages given in years)

Authors	Sample	Methods	Comparison	Main results
(Mechelli et al., 2004)	25 English-Italian early bilinguals (L2 AoA<5) 33 English-Italian late bilinguals (10<L2 AoA<15) 25 English monolinguals.	VBM analysis (density) Low-resolution method for registration (in SPM versions older than SPM5)	Early bilinguals vs. monolinguals Late bilinguals vs. monolinguals	**Left IFG**: all bilinguals > monolinguals (FWE correction at voxel-level)
(Ressel et al., 2012)	22 Catalan-Spanish bilinguals (L2 AoA<7; mean age, 23.1). 22 Spanish monolinguals (mean age, 21.5). Matched for gender.	1.5T scanner, voxel size:1mm^3 VBM & ROI analysis, Modulated images (volume), Standard Unified segmentation (in SPM8), DARTEL for own template creation, High-resolution registration, 8mm (FWHM) WM+GM as covariate.	Early bilinguals vs. monolinguals	**No significant differences** (FWE correction at voxel-level) **Left Heschl**: bilinguals > monolinguals (SVC of the FWE) **Left/right Heschl**: bilinguals > monolinguals (ROI approach)

(Zou et al., 2012)	14 Chinese-CSL bimodal bilinguals (mean L2 AoA, 19; mean age, 49) 13 Chinese monolinguals (mean age, 48)	3T scanner, voxel size:1.3x1.0x1.3mm^3 VBM analysis, Modulated images (volume), Optimized VBM5 protocol (SPM5), Low-resolution registration.	bimodal bilinguals vs. monolinguals	**Left Caudate**: bimodal bilinguals > monolinguals (SVC of the FWE)
(Gold et al., 2013)	20 English-variable_L2 bilinguals (L2 AoA<10; mean age, 63.9) 20 English monolinguals (mean age, 64.4) Matched for gender	3T scanner, voxel size:1mm^3, VBM analysis, Modulated images (volume), Standard unified segmentation (SPM8), Own template creation, High-resolution registration, 8mm (FWHM), TIV as covariate.	bilinguals vs. monolinguals	**No significant differences** (FWE correction at voxel-level)

(Grogan et al., 2012)	31 multilinguals (variable_L1, English as L2, variable_L3; mean age, 26.7) 30 bilinguals (variable_L1, English as L2; mean age, 26.7) L2 AoA balanced between groups	1.5T scanner, voxel size:1mm^3, VBM analysis, Modulated /unmodulated images (volume/density), Standard unified segmentation (SPM5), Low-resolution registration, 8mm (FWHM), Age as covariate.	multilinguals vs. bilinguals	**No significant differences** (FWE correction at voxel-level) **Right IPL**: multilinguals > bilinguals (SVC of the FWE) (just in density images)
(Abutalebi et al., 2013)	14 German-Italian-English multilinguals (L2 AoA<5; L3 AoA>10 years) 14 Italian monolinguals Groups matched in age (mean age, 23.5), all females	3T scanner, voxel size:1mm^3, VBM analysis, Modulated images (volume), Optimized VBM5 protocol, Low-resolution registration, 4mm (FWHM), TIV as covariate.	multilinguals vs. monolinguals	**Left putamen**: multilinguals > monolinguals (SVC of the FWE)
(Pliatsikas et	17 Greek-English bilinguals	3T scanner, voxel size:1mm^3,	Late bilinguals vs. monolinguals	**Right/left cerebellum**: bilinguals > monolinguals

al., 2014)	(L2 AoA>6; mean L2 AoA, 7.7; mean age, 27.5). 22 English monolinguals (mean age, 24.5)	VBM analysis, Modulated images (volume), FSL-VBM protocol, Own template creation, Medium-resolution registration, 3mm (sigma), Age and gender as covariates.		(TFCE correction)
(Abutalebi et al., 2014)	12 Cantonese-English bilinguals 11 Cantonese-Mandarin bilinguals (mean AoA, 18.87; mean age, 62.17) 23 Italian monolinguals (mean age, 61.9).	3T scanner, voxel size:1mm^3, VBM & ROI analysis, Modulated images (volume), Optimized VBM8 protocol, Low-resolution registration, East Asian brain ICBM template for bilinguals, European brain ICBM template for monolinguals,	Late bilinguals vs. monolinguals	**Left aITG**: bilinguals > monolinguals (FDR correction at cluster-level) **Left/right OFC, TmP**: bilinguals > monolinguals (ROI approach)

		DARTEL for registration, High-resolution registration, 8mm (FWHM) Sex, TIV, education level and age as covariates.		
(Klein et al., 2014)	12 simultaneous bilinguals (L2 AoA<3; mean age, 23), 25 early bilinguals (4<L2 AoA<7; mean age, 26), 29 late bilinguals (8<L2 AoA<13; mean age, 28), All French-English bilinguals 22 English monolinguals	1.5T scanner, voxel size: 1mm^3, Cortical thickness analysis, Vertex-based approach, CIVET processing pipeline	Simultaneous bilinguals vs. monolinguals Early bilinguals vs. monolinguals Late bilinguals vs. monolinguals	**Left IFGTr, left IFGOr**: early, late bilinguals > monolinguals **Right IFGOr**: early bilinguals < monolinguals. late bilinguals < monolinguals, simultaneous & early bilinguals. (FDR correction at whole-brain)
(Abutalebi, Guidi, et al., 2015)	11 Cantonese-English bilinguals 8 Cantonese-	3T scanner, voxel size:1mm^3, VBM analysis,	Late bilinguals vs. monolinguals	**Left/Right ACC**: bilinguals > monolinguals (FDR correction at

	Mandarin bilinguals (mean AoA, 12.68; mean age, 61.68) 19 Italian monolinguals (mean age, 60.93).	Modulated images (volume), Optimized VBM8 protocol, Low-resolution registration, East Asian brain ICBM template for bilinguals, European brain ICBM template for monolinguals, DARTEL for registration, High-resolution registration, 8mm (FWHM)		cluster-level)
(Abutalebi, Canini, et al., 2015)	16 Cantonese-English bilinguals 14 Cantonese-Mandarin bilinguals (mean AoA, 18.27; mean age, 63.2) 30 Italian monolinguals (mean age, 61.85).	3T scanner, voxel size:1mm^3, ROI analysis, Modulated images (volume), Optimized VBM8 protocol, Low-resolution registration,	Late bilinguals vs. monolinguals	**Left/Right IPL**: bilinguals > monolinguals (ROI approach)

		East Asian brain ICBM template for bilinguals, European brain ICBM template for monolinguals, DARTEL for registration, High-resolution registration		
(Olsen et al., 2015)	14 English-variable_L2 bilinguals (L2 AoA<11) 14 English monolinguals (mean age, 70.6).	3T scanner, voxel size:1mm^3, ROI analysis, (volume) linear and nonlinear registration (ANT algorithm),	Late bilinguals vs. monolinguals	**No significant differences** (Explore GM volume from temporal, parietal, frontal and occipital lobe)
		Cortical thickness analysis, ROI approach, Freesurfer processing pipeline.		**No significant differences** (Explore CT from entorhinal cortex, hippocampus and temporal pole)

Abbreviations: aITG=anterior inferior temporal gyrus; AoA=age of acquisition; FDR=false discovery rate; FWE=family wise error; FWHM=full-width at half-maximum; GM=grey matter; IFGOr=pars orbitalis, inferior frontal gyrus; IFGTr=pars triangularis, inferior frontal gyrus;

IPL=inferior parietal lobule; OFC=orbito-frontal cortex; ROI=region of interest**;** STG=superior temporal gyrus; SVC=small volume correction; TFCE=threshold free cluster enhancement; TIV=total intracranial volume; TmP=temporal pole; VBM=voxel-based morphometry; WM=white matter; ACC=anterior cingulate cortex; CT=cortical thickness.

Table 2: Cross-sectional WM studies related to bilingualism (ages given in years)

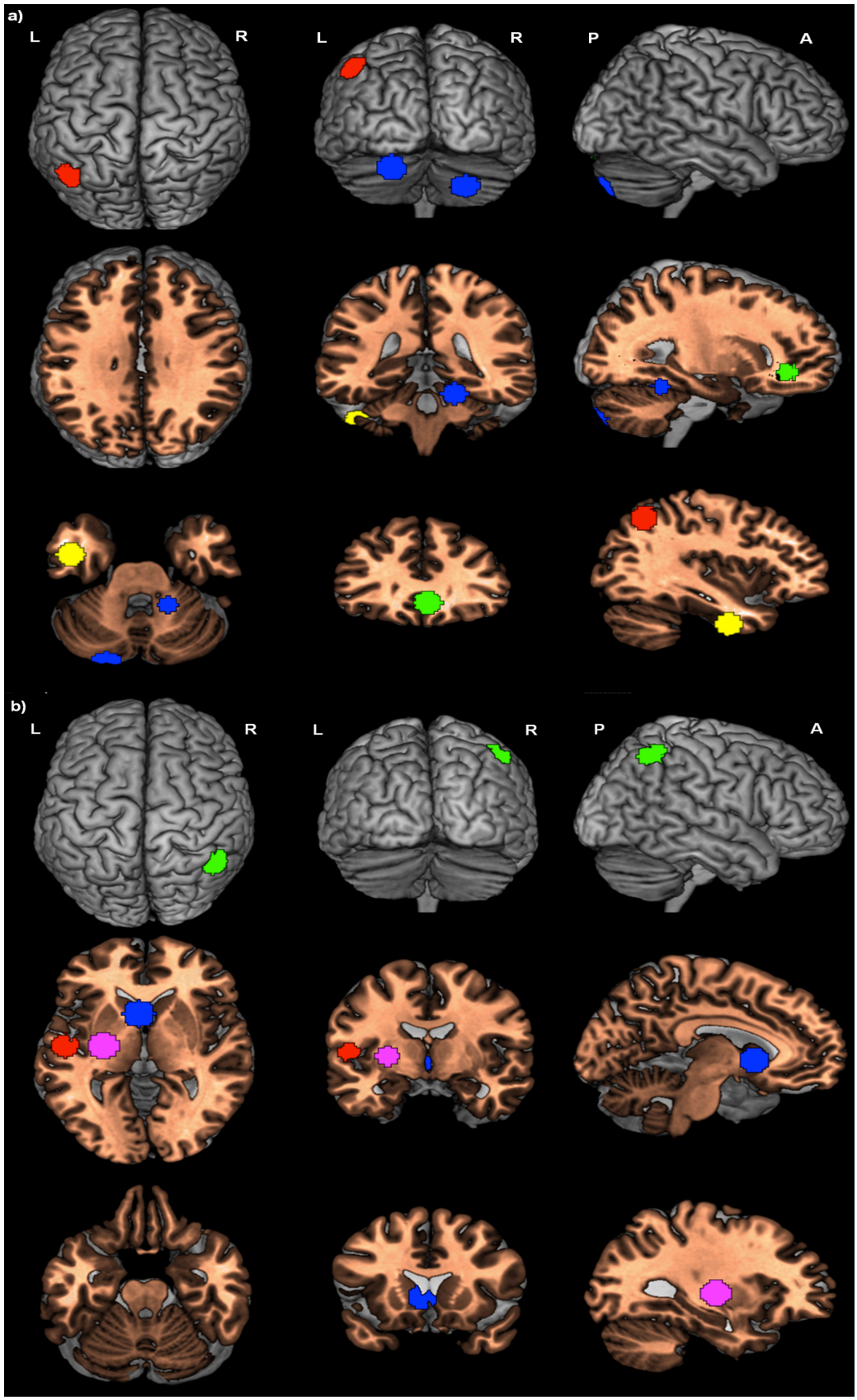
Authors	Sample	Methods	Comparison	Main results
(Mechelli et al., 2004)	25 English-Italian early bilinguals (L2 AoA<5) 33 English-Italian late bilinguals (10<L2 AoA<15) 25 English monolinguals.	VBM analysis Unmodulated images (density)	Early bilinguals vs. monolinguals Late bilinguals vs. monolinguals	**No significant differences** (FWE correction at voxel-level)
(Ressel et al., 2012)	22 Catalan-Spanish bilinguals (L2 AoA<7, mean age, 23.1). 22 Spanish monolinguals (mean age, 21.5). Matched for gender.	1.5T scanner, voxel size:1mm^3 VBM & ROI analysis, Modulated images (volume), Standard Unified segmentation (in SPM8), DARTEL for own template creation, High-resolution registration, 8mm (FWHM) WM+GM as covariate.	Early bilinguals vs. Monolinguals	**No significant differences** (FWE correction at voxel-level)

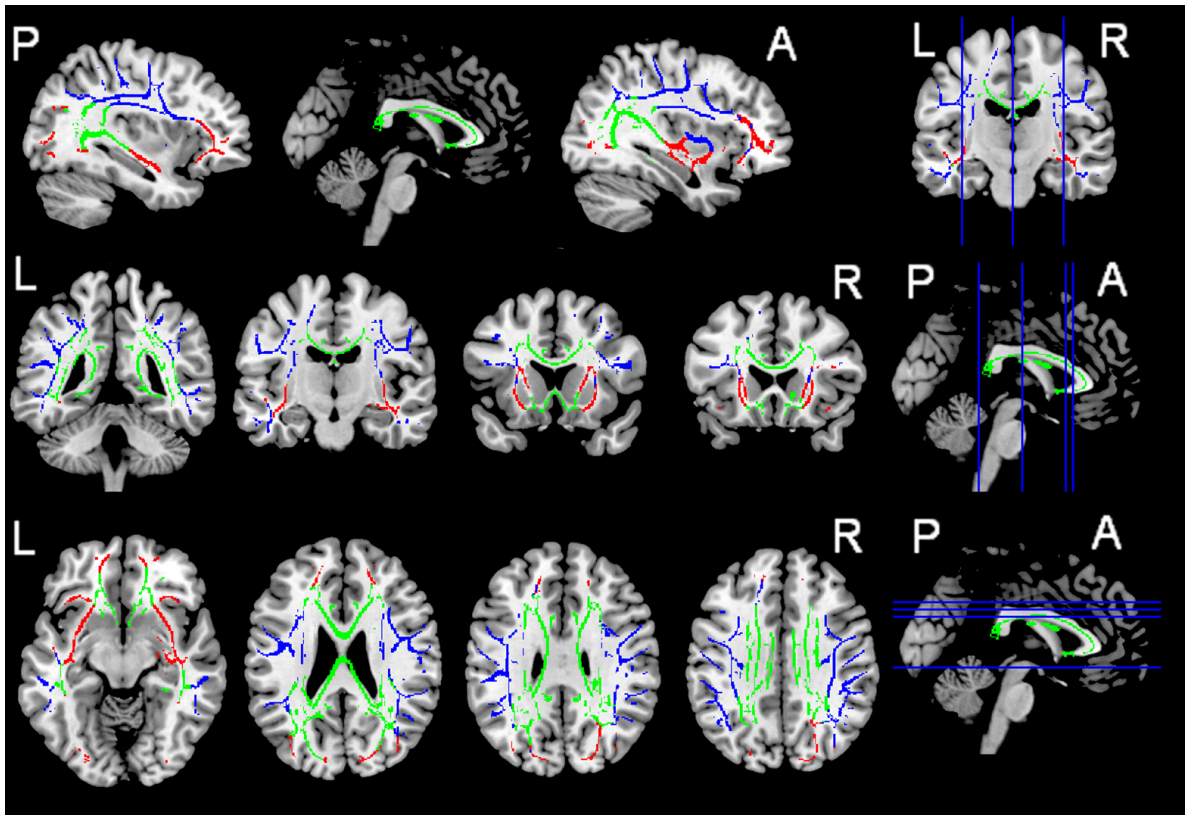
(Luk et al., 2011)	14 English-variable_L2 bilinguals (L2 AoA<11yo) 14 English monolinguals (mean age, 70.5)	3T, 30 directions, 5mm slice thickness, TBSS protocol FA, RD, AD. Sample-specific target image for registration, Medium-resolution registration	Bilinguals vs. Monolinguals	**FA in CC, SLF, Right IFOF and uncinate:** Bilinguals>Monolinguals **RD in CC:** Monolinguals>Bilinguals
(Mohades et al., 2012)	15 Dutch-variable_L2 simultaneous bilinguals (L2 AoA<3, mean age, 9.3) 15 Dutch-variable_L2 sequential bilinguals (L2 AoA>3, mean age, 9.7) 10 Dutch monolinguals (mean age, 9.6)	3T scanner, 15 directions, voxel resolution: $1.75x1.75x2mm^3$, TOI analysis, FA mean values	Simultaneous bilinguals vs. Monolinguals Simultaneous vs. Sequential bilinguals Sequential bilinguals vs. Monolinguals	**Mean FA in both IFOF:** Simultaneous>Sequential bilinguals>Monolinguals **Mean FA in anterior CC to orbital lobe tracts:** Monolinguals>Simultaneous bilinguals (Bonferroni correction)
(Gold et al.,	20 English-variable_L2	3T scanner, 36 directions, voxel	Bilinguals vs. Monolinguals	**FA in both ILF/IFOF, fornix, CC**

2013)	(L2 AoA<10; mean age, 63.9) 20 English monolinguals (mean age, 64.4) Matched for gender	resolution: 1.75x1.5x3mm^3, TBSS protocol FA, RD, AD, MD 5000 permutations		Monolinguals>Bilinguals **RD in IFOF, CC:** Bilinguals>Monolinguals (TFCE correction)
(Cummine & Boliek, 2013)	13 Chinese-English bilinguals (L2 AoA>5, mean age, 24.2) 11 English monolinguals (mean age, 28.5)	1.5T, 12 directions, 4mm slice thickness, TBSS protocol FA, MD	Bilinguals vs. Monolinguals	**Right IFOF & Anterior Thalamic Radiation** ***(Right superior portion & inferior portion bilaterally):*** Monolinguals>Bilinguals
(Pliatsikas et al., 2015)	20 variable_L1-English (L2 AoA<10.15; mean age, 31.85; mean immersion, 91 months) 20 English monolinguals (mean age, 28.16)	3T, 30 directions, 2mm slice thickness, TBSS protocol FA	Sequential late learners vs. monolinguals	**Bilaterally CC, IFOF, Uncinate and SLF:** Bilinguals >monolinguals
(Olsen et al., 2015)	14 English-variable_L2 bilinguals (L2	3T scanner, voxel size:1mm^3, ROI analysis,	Late bilinguals vs. monolinguals	**Mean volume in frontal lobe:** bilinguals > monolinguals

	AoA<11yo) 14 English monolinguals (mean age, 70.6). (Same sample as Luk et al., 2011)	(volume) linear and nonlinear registration (ANT algorithm)		(ROI approach)
(Mohades et al., 2015)	14 Dutch-variable_L2 simultaneous bilinguals (L2 AoA<3, mean age, 11.4) 16 Dutch-variable L2 sequential bilinguals (L2 AoA>3, mean age, 11.33) 10 Dutch monolinguals (mean age, 11.8) (Same sample as Mohades et al., 2011, but 2 years later)	3T scanner, 15 directions, voxel resolution: 1.75x1.75x2mm^3, TOI analysis, FA mean values	Simultaneous bilinguals vs. Monolinguals Simultaneous vs. Sequential bilinguals Sequential bilinguals vs. Monolinguals	**Mean FA in left IFOF:** Simultaneous>Sequential bilinguals>Monolinguals (TOI approach, Bonferroni correction)

Abbreviations: AD=axial diffusivity; AoA=age of acquisition; CC=corpus callosum; FA=fractional ansiotropy; FWE=family wise error; FWHM=full-width at half-maximum; GM=grey matter; IFOF=inferior frontal-occipital fasciculus; ILF=inferior longitudinal fascicule; MD=mean diffusivity; RD=radial diffusivity; ROI=region of interest; SLF=superior longitudinal fasciculi; TBSS=tract-based spatial statistic; TFCE=threshold free cluster enhancement; TOI=tract of interest; VBM=voxel-based morphometry; WM=white matter.


a)
L
R
L
R
P
A
b)
L
R
L
R
P
A


P
A
L
R
L
R
P
A
L
R
P
A