Referring expressions and executive functions in bilingualism

Antonella Sorace
University of Edinburgh

Address for correspondence:
University of Edinburgh
School of Philosophy, Psychology and Language Sciences
3 Charles Street
Edinburgh EH8 9AD
Scotland, UK
email: antonella@ling.ed.ac.uk

Word count: 4554
Abstract
Recent research has shown that the bilingual experience has positive effects on non-linguistic cognition (Bialystok 2009; Costa and Sebastian-Gallés 2014) but also negative effects on language, for example on vocabulary size and lexical fluency (Pearson et al. 1993). While most of the linguistic ‘disadvantages’ of bilingualism have been discussed in the lexical domain, this question is scaled up here to the sentence level and a novel theoretical framework is proposed which explicitly connects psychological and linguistic research. It is suggested that the bilingual experience may (a) affect the reciprocal interactions between language and general cognition, and (b) modulate the relation between components of executive functions. These effects may in turn influence the processing of particular linguistic structures, such as anaphoric expressions, and lead to bilingual-monolingual differences that could be regarded as ‘disadvantages’ but are in fact the result of normal adaptive changes due to the bilingual experience.

Introduction
One of the most conspicuous – and most controversial - findings from research on language and cognition in bilinguals is that knowing more than one language brings a number of beneficial changes across the lifespan. Children who know more than one language have a spontaneous understanding of language structure and therefore an enhanced ability to learn new languages, as well as an earlier grasp of some essential background components of literacy, such as the invariance of print meaning and its symbolic function (Bialystok 2002). Moreover, bilingualism gives children and adults advantages in tasks that involve cognitive flexibility and the control of attention: bilinguals seem to be better at selectively paying attention, at inhibiting irrelevant information, and at switching between alternative solutions to a problem (Bialystok & Martin 2004; Bialystok 2009; see Vega-Mendoza et al 2015 on language learning in young adults). Importantly, these benefits do not appear across the board: for example, bilinguals do not seem to have an advantage over monolinguals with respect to functions that depend on the way knowledge is represented, such as encoding problems or drawing logical inferences. At the root of these cognitive effects is the bilingual’s constant experience of having two languages simultaneously active and inhibiting one when the other is used (Costa et al. 2008; Green 1998), which enhances executive control in other domains.
Many questions remain open: for example, whether bilinguals are more efficient at inhibition of irrelevant information, or whether they have an enhanced ability to selectively activate relevant information (Costa, Santesteban and Ivanova 2006); whether they acquire a range of more subtle advantages, such as the ability to ‘modulate’ executive function according to the type of task they engage in (Blumenfeld and Marian 2011); whether the overlap between inhibitory control for language overlaps completely or only partially with inhibitory control in non-linguistic cognition (see Calabria et al 2012 for evidence of qualitative differences between the two); whether it would ultimately be preferable to assume a more ‘unified’ account that takes into account the whole attentional system rather than isolated components (Bialystok 2015); and whether the source of the bilingual advantage may lie in post-conflict disengagement of attention (Mishra et al 2012; Grundy and Shedden 2014). The most recent debate has centered in particular on the replicability of the bilingual advantage, which a number of studies have failed to find (Paap and Greenberg 2013; Duñabeitia et al. 2014; see Valian 2015 for an overview). Some researchers interpret these null results as questioning the validity of previous results showing a bilingual advantage (de Bruin et al., 2015). Others (Baum and Titone 2014; Kroll and Bialystok 2013) view the failure to replicate in some studies as a normal manifestation of variation due to interactions with poorly understood factors (age at testing, distance between the two languages, patterns of bilingual language use, education levels, societal attitudes, etc.), and therefore as an incentive to carry out more research that compares different types of bilingualism and bilingual settings. In any case it should be borne in mind that the effect of bilingualism on executive functions is only one of many aspects of the bilingual experience that have been studied extensively.

The discussion in the literature about the effects of bilingualism on cognitive control and executive functions has so far been largely confined to cognitive psychology and based on the findings from psychological experiments. An unexplored side of the question is whether the changes in cognitive control due to the bilingual experience can selectively affect particular aspects of language processing. Individual differences in cognitive control abilities have been found in previous studies to influence sentence-level processing abilities (Vuong & Martin 2014; Novick et al 2014; Teubner et al 2016). It is unclear, however, whether these effects are more likely to be
seen for certain types of structures. It is possible, for example, that individual differences in cognitive control abilities are more visible in processing structures that require probabilistic rather than categorical operations (Nieuwland and Van Berkum 2006). Anaphoric referential expressions, such as pronominal forms, are an example of a structure that involves probabilistic processing. Subject pronouns in null-subject languages such as Italian, Spanish and Greek are syntactically licensed but their distribution is governed by discourse-pragmatic factors (Rizzi 1982; Grimshaw and Samek-Lodovici 1998). The interpretation and production of pronouns are therefore dependent on the on-line efficient computation of these factors in real-time processing. In what follows, a new framework is described that explicitly integrates research on executive functions in bilingualism with well-documented phenomena from linguistic research on subject pronouns in late adult bilingualism. It will be suggested that connecting the two research strands can benefit our understanding of late bilingualism and provide a novel perspective on the study of the adaptability of executive functions over the lifespan.

**Anaphoric expressions in bilinguals**

Adult late bilinguals are speakers who have learned a second language after the age of 15 and have reached a high proficiency level in this language. A robust finding that has emerged from research is that monolingual and adult late bilingual speakers of Italian (and other null subject languages, such as Greek and Spanish) diverge in their production and comprehension of pronominal subjects: this divergence is manifested in the greater variability shown by bilingual speakers, regardless of whether Italian is their native (L1) or their non-native (L2) language (Belletti, Bennati & Sorace 2007; Sorace 2003, 2005, 2006a,b; Sorace & Filiaci 2006; Sorace 2011; see also Tsimpi & Sorace 2006 on L2 Greek; Tsimpli et al 2004 on Italian and Greek speakers who are in a situation of attrition due to prolonged exposure to a second language, henceforth ‘L1 attriters’; Chamorro et al 2015 on Spanish L1 attriters). In both L2 speakers and L1 attriters, variability results in the overextension of the scope of the overt subject pronoun to contents in which a null pronoun would be expected, but not vice versa. The magnitude of this overextension, however, is greater in L2 speakers than in L1 attriters. Let us illustrate the phenomenon in production and comprehension.

In production, bilingual Italian speakers are more likely to optionally utter sentences
such as (1b), with a ‘redundant’ overt pronoun, whereas a monolingual Italian speaker would produce (1c) with a null pronoun.

(1) a. Perchè Maria è arrivata così tardi?
   why Maria is arrived so late
   ‘Why did Maria arrive so late?’

   b. Perchè lei si era addormentata
      because she herself was asleep
      ‘Because she fell asleep’

   c. Perchè Ø si era addormentata
      because Ø herself was asleep
      ‘Because she fell asleep’

In contrast, errors involving null pronouns in inappropriate contexts are unattested; for example, pronouns are not omitted when a less salient referent is referred to (as in 2b), or when the sentence is explicitly contrastive (as in 3b).

(2) a. Perchè Maria ha chiamato Paolo?
   why Maria has called Paolo
   ‘Why did Maria call Paolo?’

   b. *Perchè Ø voleva vederla (Ø = lui/Paolo)
      because Ø wanted to see-her
      ‘Because he wanted to see her’

(3) a. Maria ha detto che passava a prendere Paolo?
   Maria has said that was going to pick up Paolo?
   ‘Did Maria say that she would pick up Paolo?’

   b. *No, Ø ha detto che passava a prendere lei (Ø = lui/Paolo)
      No, Ø has said that was going to pick up her
      No, he said that she would pick up her’

The greater variability of overt pronouns is attested not only in bilingual speakers’ production, but also in their interpretation of pronominal subjects. This is particularly clear with respect to intersentential anaphora involving two clauses, one including
two equally plausible antecedents and one containing an overt pronoun. In forward anaphora (where the antecedents precede the pronoun, as in Table 1), bilingual Italian speakers often interpret the overt pronominal subject of the embedded clause as coreferential with the lexical subject of the main clause (*Mario*), whereas monolingual Italian speakers prefer to interpret the overt pronoun in this context as referring to the complement (*suo fratello*, ‘his brother’). In contrast, the null subject pronoun is preferentially interpreted as referring to the subject antecedent by both monolingual and bilingual speakers (Sorace and Filiaci 2006; Tsimpī et al 2004)

<table>
<thead>
<tr>
<th>MONOLINGUAL ITALIAN</th>
<th>BILINGUAL ITALIAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mario non vede <em>suo fratello</em> da quando <em>lui</em> è partito</td>
<td>Mario non vede suo fratello da quando <em>lui</em> è partito</td>
</tr>
<tr>
<td><em>Mario hasn’t seen his brother since he left</em></td>
<td><em>Mario hasn’t seen his brother since he left</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MONOLINGUAL ITALIAN</th>
<th>BILINGUAL ITALIAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mario non vede suo fratello da quando <em>o</em> è partito</td>
<td>Mario non vede suo fratello da quando <em>o</em> è partito</td>
</tr>
<tr>
<td><em>Mario hasn’t seen his brother since he left</em></td>
<td><em>Mario hasn’t seen his brother since he left</em></td>
</tr>
</tbody>
</table>

Table 1: Differences between monolingual and bilingual interpretations of Italian anaphoric forms

In backward anaphora (i.e. when the clause including the pronoun precedes the clause containing the referents), monolingual speakers typically interpret the overt subject as referring to the object, as in (4a) or to an extralinguistic referent (Kraš et al 2014; Sorace & Filiaci 2006); bilinguals, on the other hand, are more likely than monolinguals to establish a dependency between the overt pronoun and the matrix subject, as in 4b.

(4) a. Quando leiₖ era in città, Paolaₖ è andata a trovare Mariaₖ.
    b. Quando leiᵢᵢ era in città, Paolaᵢᵢ è andata a trovare Mariaᵢᵢ "When she was in town Paola is gone to visit Maria"

‘When she was in town, Paola went to visit Maria’
How can these patterns be explained? The remainder of this paper focuses on possible accounts based on different linguistic or cognitive factors involved in pronominal use, some of which have been researched in the literature and some that are currently unexplored. The brief descriptions show that pronominal use entails a close interplay of language and general cognition, and executive functions are likely to play a crucial role in the computation of anaphoric dependencies: for this reason, purely linguistic explanations have limited scope and need to include consideration of the type of cognitive control underlying particular linguistic operations. In particular, the attested bilingual behaviour may involve a trade-off between particular aspects of cognitive control which has not so far been researched in connection with pronominal use.

Potential explanations

Cross-linguistic influence is a type of explanation for these phenomena that has been frequently proposed for null subject languages: bilingual speakers’ knowledge representations in each language are influenced by the other language (which is English in many of the studies cited – see Sorace and Filiaci 2006; Tsimpli et al 2004). In both L2 speakers and L1 attriters, English as the language that has the least restrictive anaphoric system (no pronominal choice dependent on pragmatic factors) affects the other, regardless of whether it is L1 or L2. In L1 attriters, this influence takes the form of a neutralization of L1 pragmatic distinctions towards the less restrictive L2 system. In L2 speakers, it takes the form of a neutralization of L2 distinctions towards the less restrictive L1 system.

However, this account is insufficient to explain why the overextension of overt pronouns is also attested in adult bilingual speakers of two null subject languages of the same type (Italian-Spanish, Greek-Spanish, Spanish-Portuguese; e.g. Bini 1993; Malgaza & Bel 2006; Lozano 2007; Mendes & Iribarren 2007; de Prada 2009). The irrelevance of typological similarity strongly suggests that language interference cannot be the only cause of this phenomenon. A similar indication comes from developmental patterns of asymmetric extension of overt pronominal subjects in bilingual L1 acquisition (Serratrice et al. 2009; Sorace, Serratrice, Filiaci, & Baldo (2009); Sorace & Serratrice 2009; Serratrice et al, 2012). Sorace et al (2009)
conducted a large-scale study in which they compared two groups of school-age bilingual children acquiring two different combinations of languages; Italian-English (in which only one language allows null subjects) and Italian-Spanish (in which both languages allow null subjects). Elicited preference experiments showed that both child bilingual groups accepted significantly more overt subjects referring to topic antecedents (as in Paperino, ha detto che lui, è caduto ‘Donald Duck, said that he, fell’) than monolingual children, regardless of language combination. Moreover, the younger monolinguals also did this significantly more often than the adult controls, indicating that these aspects of the syntax-pragmatics interface are acquired late (Sorace & Serratrice 2009).

A different type of explanation focuses on **real-time processing**, since the use of pronominal forms requires the efficient integration and coordination of grammatical and pragmatic information in real time (Sorace 2011, 2012). In natural interaction, bilingual speakers have to be able to rapidly exclude irrelevant pronoun-antecedent mappings, integrate changing information from the context and from the assessment of the interlocutor’s knowledge state, and update the representation of the situation accordingly (see Brown-Schmidt 2009). The efficiency of these operations may be variable for both monolingual and bilingual speakers.

Indeed, psycholinguistic research on anaphora resolution in monolingual native speakers of null-subject languages lends support to this argument. Carminati (2002, 2005) provides experimental evidence that null and overt pronouns in Italian have distinct and complementary functions, manifested in their distinct biases for antecedents in different syntactic positions. Null pronouns have a strong bias towards an antecedent in Spec IP (normally – but not exclusively – the subject), whereas overt pronouns prefer an antecedent in positions lower in the phrase structure (normally – but not exclusively – a complement): this is referred to as the ‘Position of Antecedent Strategy’ (henceforth PAS). The PAS, for Carminati, is a highly efficient processing principle that belongs to the interface between syntax and discourse: not only is there a reliable correspondence between the structural position Spec IP and the notion of topic, but also pragmatic principles are the core of antecedent preferences. So, for example, using an overt pronoun to refer to a topic antecedent would represent a violation of Grice’s maxim of quantity, because since another form – the null pronoun
- is available for the same purpose, the comprehender assumes that it should have been used instead. Crucially, however, there is a difference between null and overt pronouns with respect to the strength of the PAS. Carminati’s experimental data indicate that while the preference of the null pronoun for subject antecedents is very consistent, antecedent preferences for the overt pronoun are more flexible: a weaker processing cost may be incurred if an overt pronoun takes a subject antecedent than if a null pronoun takes a non-subject antecedent. The antecedent preferences of overt pronouns appear to be sensitive to contextual factors: monolingual speakers are more tolerant of PAS violations in unambiguous sentences, in which the potential for miscommunication is low. It appears, therefore, that monolingual speakers may be occasionally unable or unwilling to engage in full processing when they know that the context is sufficiently unambiguous, as in (5b), in which there is only one referent, or (5c), in which the pronoun agrees in number with only one of the two antecedents; in these cases they may produce a sentence with an unnecessary, or redundant overt pronoun which does not impair antecedent assignment in comprehension. An overt pronoun would be much less likely to be produced in the ambiguous context of (5a).

(5) a. Paola, passava molto tempo con Luisa quando lei era in vacanza

   ‘Paola used to spend a lot of time with Luisa when she was on holiday’.

b. Giorgio ha detto che lui non vota alle prossime elezioni

   ‘Giorgio said that he will not vote at the next election’

c. Quando Carlo ha visto i suoi amici, lui era molto contento

   ‘When Carlo saw his friends he was very happy’

Thus, overt pronouns may be used inappropriately when the speaker does not pay enough attention to encoding the utterance from the comprehender’s perspective, or is otherwise unable to do so when, for example, the processor is overloaded: in this case, the PAS is relaxed, although comprehensibility is not compromised. It is plausible to think that bilingual speakers, whose processing resources are more taxed, may resort to relaxing the PAS in a wider range of contexts. The overt pronoun may therefore be a default form used to relieve processing demands when these become temporarily...
unmanageable. If these assumptions are correct, one would expect that these patterns of pronoun overgeneralization in Italian should be produced not only by bilinguals who speak English as one of their languages, but also by bilingual speakers of different language pairs, including languages that have a similar pronominal system to that of Italian: exactly what emerges from the studies just reviewed. In other words, the difference between monolinguals and bilinguals may be more quantitative than qualitative.

A similar conclusion can be reached on the basis of other models suggesting that both monolingual and bilingual speakers may experience fluctuations in the processes of integration and updating of contextual cues that signal changes in pronoun-context mappings. According to the two-step model of reference tracking developed by Hendriks, Koster & Hoeks (2014), choosing a referring expression in production consists of (a) first selecting the most reduced (default) form, and (b) next, selecting a form that can be best understood by the listener if adjustment is needed. While maintaining reference to salient topical antecedents should not be a problem, the production of more explicit forms to signal reference to less salient referents is costly: it requires mentalizing about the listener’s potential interpretation, inhibiting the less informative pronominal forms, and updating the mental representation of the situation. What is interesting from this perspective is that bilinguals are over-explicit: they produce fewer reduced forms. This suggests that they do not find switching reference problematic, but rather that they may have a higher threshold for deciding that a reduced form is sufficiently unambiguous, possibly as a consequence of enhanced perspective-taking abilities. In comprehending referential forms, the interpretation of pronouns may initially be based on a default mapping to the most prominent referent, which is unproblematic in topic maintenance contexts. However, hearing a less reduced form signals a shift of reference to a less prominent referent from the perspective of the speaker. Interpreting such forms again involves mentalizing about the speaker’s intention and updating the representation of the situation accordingly. Bilinguals may not be consistently successful at these operations, depending on the cognitive resources that they can recruit at any one time.

**Competition for resources and cognitive load** are in fact critical factors in the coordination of constantly changing pronoun-context mappings in the real-time use of
anaphoric expressions. Since these processes are consuming in terms of cognitive resources, one would expect inconsistency and occasional ‘discoordination’ of pronominal use in populations that are more sensitive to cognitive load. Discoordination in pronominal reference has in fact been attested in ageing speakers (Titone et al 2000), schizophrenic patients (Phillips & Silverstein 2003), and children with autism (Arnold, Bennetto & Diehl 2009). Bilinguals need to exercise executive control to avoid interference from the unwanted language. Suppose that anaphoric dependencies partly draw on the same pool of attentional resources used to keep the two languages separate: this creates a competition for resources when bilinguals engage in linguistic tasks that are sensitive to cognitive load, which may impact on different aspects of the task. In the case of anaphoric dependencies, the assessment of the interlocutor’s knowledge state and of the relative accessibility of referent may (inconsistently) exceed the speaker’s resources. As Keysar, Lin & Barr (2003) argue, adult speakers do not reliably consider what the interlocutor knows in their initial encoding of referential expressions, and resources are needed to recover from initially ‘egocentric’ computations. Asymmetric inhibition effects (Meuter & Allport 1999) may account for the different extent to which overt pronouns are overextended by L1 attriters and L2 speakers: in L2 speakers, the unwanted language is the (still dominant) L1, which requires more resources to be inhibited; in L1 attriters, in contrast, the unwanted language is the (less dominant) L2, which requires fewer resources to be inhibited.

A trade-off between inhibitory control and integration/updating is an alternative and so far unexplored account. Increased inhibitory control and less efficient integration/updating ability may be in a trade-off relationship, in a similar way to the relationship between inhibitory control and negative priming (Treccani et al 2009). Integration of cues that signal switching to a different interpretation, for example, requires “disengagement” of inhibition (Blumenfeld & Marian 2011). The two components have been found to be dissociated in several impaired and typical populations (Titone et al 2000; Phillips & Silverstein 2003; Watson et al 2012; Arnold, Bennetto & Diehl 2009). If the ability to integrate and update is in a trade-off relation with inhibitory control (see e.g. Braver 2012; Goschke and Dreisbach 2008), one might expect to see variability and inconsistency in reference tracking which depends on the relative strength of one or the other aspects of executive function in
particular groups or on moment-by-moment fluctuations in attentional control within individual speakers. Recent research on the effects of bilingualism on executive functions has shifted the focus from the role of inhibitory control in conflict resolution to the ability to adjust and refocus attention in a continuously changing environment (see e.g. Mishra et al 2012 for results showing an early bilingualism advantage in this respect). Variables such as age of onset of bilingualism and/or balance between the two languages may have an influence on the way bilingual speakers deal with the trade-off tension between inhibitory control and integration/updating. It has been argued (Costa & Santesteban 2004) that there may be differences between early and late bilinguals (or between more balanced and less balanced bilinguals) with respect to the presence or the type of effects of the bilingual experience on executive function. These differences may be due to the fact that executive functions in early bilinguals develop in a way that is optimally suited to the use of two languages, whereas late bilinguals learn a second language with an attentional system that, at least initially, is optimally suited for the use of only one language. Early bilinguals who make frequent use of both their languages may be predicted to acquire the ability not only to apply inhibitory control, but also to ‘disengage’ inhibition when required by the nature of the task; disengagement of inhibition allows more flexibility in task switching and facilitates updating of the mental representation of the situation. Late bilinguals, on the other hand, may develop enhanced inhibitory control because of the need to apply more inhibition to their dominant L1 when they speak the L2, without having the long-term experience of using both languages and switching between the two. While no research has so far directly compared early and late bilinguals in post-conflict resolution tasks, recent results (Bak, Vega-Mendoza & Sorace (2014) indirectly support the hypothesis that late bilinguals may have an advantage in inhibitory control but not in task-switching and adapting attention to new conditions.

One of the experiments in this study employed three tests from the Test of Everyday Attention (TEA, Robertson et al. 1994) of increasing complexity, which measured (from least to most complex) sustained attention, selective attention and inhibition, and task switching and monitoring. The battery was administered to monolinguals, early childhood bilinguals, late childhood bilinguals, and young adulthood bilinguals, with a variety of language background and language combinations. Both early and late bilinguals outperformed monolinguals, but in different tasks: while the advantage for early bilinguals was larger for on task switching test, the advantage for adult
bilinguals was evident in the inhibition test but not in the switching test. However, much more research on bilingual and multilingual speakers of different language combination and different ages of first exposure to a second language is necessary to explore these differential effects on cognitive control. Future research will also establish whether disengagement of inhibition might be at work in the use of anaphoric expressions and whether it may be in part responsible for the different extent to which child and adult bilinguals resort to the use of overt pronouns as a default.

Conclusions
This paper has presented a theoretical exploration of variability in pronominal use - a well-attested linguistic phenomenon in bilingualism – from the point of view of cognitive control and executive functions. The proposal opens up four new ways of thinking about the relationship between executive functions and bilingualism. First, linguistic research on bilingualism can benefit from integrating findings from psychological research on executive functions, especially for structures, such as pronominal use, that involve connections between linguistic and non-linguistic factors.
Second, bilingual language behaviour beyond the lexical level may be informative about the effects of the bilingual experience on general cognition. Investigating the aspects of executive functions involved in the use of particular language structures, and at how they vary among monolinguals and bilinguals, can shed light on the precise locus of the bilingual effects on cognitive control and contribute to understanding the reasons why these effects are not consistently found in all bilingual contexts.
Third, bilingualism is likely to affect an array of components of executive functions and their relationship, rather than a single component (e.g. inhibition). The key for future research may be to focus on individual differences in modulating executive functions in a flexible way depending on particular tasks, and examine whether acquiring a second language at a different stage in life can impact the adaptability of cognitive control.
Fourth, pronominal use in bilingual speakers is not monolingual-like, in L2 or in L1, but is not radically different either: bilinguals tend to make more extensive use of an option that monolinguals also employ. Is this a ‘disadvantage’? The differences
between monolinguals and bilinguals in language processing, like the differences in general cognition, can be seen as advantageous or disadvantageous only if one takes the monolingual system as a point of reference. However, bilinguals are not the sum of two monolinguals, as Grosjean (2008) reminds us. The patterns of convergent bilingual pronominal use in L2 speakers and L1 attritors may be revealing a reconfiguration of the cognitive network that enables successful bilinguals to flexibly use more than one language. Reconfiguration of the language space may lead to convergence between L1 and L2, so that proficient bilinguals are not, and should not be expected to be, like monolinguals in either of their languages. Future interdisciplinary research is needed to understand individual differences in this domain, as well as the details of how language-specific and general cognitive factors interact across the lifespan and at different stages of bilingual development.
References


