A common selection mechanism at each linguistic level in bilingual and monolingual language production

Esti Blanco-Elorrieta\textsuperscript{a,}*, Alfonso Caramazza\textsuperscript{b}

\textsuperscript{a} Department of Psychology, Harvard University, Cambridge, MA, USA
\textsuperscript{b} Center for Mind/Brain Sciences (CIMcC), University of Trento, Trento, Italy

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Lexical access
Language selection
Selection-by-activation
Bilingual language organization
Bilingual communication
Bilingual language production

\textbf{ABSTRACT}

The primary goal of research on the functional and neural architecture of bilingualism is to elucidate how bilingual individuals’ language architecture is organized such that they can both speak in a single language without accidental insertions of the other, but also flexibly switch between their two languages if the context allows/demands them to. Here we review the principles under which any proposed architecture could operate, and present a framework where the selection mechanism for individual elements strictly operates on the basis of the highest level of activation and does not require suppressing representations in the non-target language. We specify the conjunction of parameters and factors that jointly determine these levels of activation and develop a theory of bilingual language organization that extends beyond the lexical level to other levels of representation (i.e., semantics, morphology, syntax and phonology). The proposed architecture assumes a common selection principle at each linguistic level to account for attested features of bilingual speech in, but crucially also out, of experimental settings.

1. Introduction

Since the onset of the empirical study of language and cognition, the fact that two distinct languages can cohabit in a single mind has sparked the interest of many researchers. Several decades in, there is still no agreed-upon answer to the central question: How do bilinguals manage to speak in the language they intend to, without constant and unwanted insertions from their other language? So far, attempts to answer this question have focused on single linguistic levels; mostly the lexical level.° No agreed-upon answer to the central question: How do bilinguals manage to speak in the language they intend to, without constant and unwanted insertions from their other language? So far, attempts to answer this question have focused on single linguistic levels; mostly the lexical level.° The fact that two distinct languages can cohabit in a single mind has sparked the interest of many researchers. Several decades in, there is still no agreed-upon answer to the central question: How do bilinguals manage to speak in the language they intend to, without constant and unwanted insertions from their other language? So far, attempts to answer this question have focused on single linguistic levels; mostly the lexical level.° However, to characterize successful bilingual communication, one has to propose principles that can generalize across linguistic levels to construct a cohesive language system. Here we explore broadly the principles on which the mechanism allowing for bilingual communication may work and decide on a set of parameters that can govern language selection at every linguistic level, leading to a bilingual language architecture that covers the whole language system. We will additionally examine the evidence postulated to support previous models and we will account for it within our proposed framework, keeping the theoretical assumptions for the lexical selection mechanism in place.

2. Operational principles

In what follows, we review the proposals that could in principle constitute the basis for bilingual language production. This includes theoretical possibilities regardless of the current level of empirical support for them, as we see value in exploring all the possible options for each of the parameters that needs to be defined in order to propose a bilingual language production framework.

The first assumption required in any model of bilingual language production is a principled representation of the distinction of the two languages. In order to have the choice of speaking one language or the other, there needs to be a way to identify which language a given element belongs to. The first possibility suggested by researchers was one where the lexicon was organized in “boxes”, with all the items that belong to each language being stored in separate boxes with a switch mechanism for determining whether the search for words was to be conducted in one box or the other (Macnamara, 1967). This architecture
essentially postulates two independent parallel organizations for each language. In this scenario, Language as a feature is fundamentally different than other types of linguistic features associated with lexical or syntactic elements such as register, dialect, etc. The latter features are selected via processes within each language that are different from the language selection mechanism. On this view, the bilingual system necessitates an additional mechanism to deal with language representation that is absent for monolingual lexical organization.

An alternative possibility for the representation of language membership is via a direct link between each element at each linguistic level and a Language node. Under this system, each item in the lexicon could have a connection to a language node, just as it can have a connection to a noun/verb node, feminine/masculine node, etc. One could imagine an architecture to be a mere extension of the one in place to represent an item belonging to a given register, dialect, etc., with the only addition to the structure in bilinguals being an extra link to a language node.

Once language membership has been established, the next principle that needs to be articulated are the rules that governs activation flow. If a bilingual wants to speak a given language, do all stored elements, even the ones in the non-intended language, receive activation? Or could a system be developed such that only the words in the relevant language will be activated? A structure that would allow for only words in the target language would independently reach elements at each linguistic level, and while it could boost the activation of an element of one language to selection threshold for an element of that same language at all levels (e.g., the lexical level), it would not necessarily boost activation to selection threshold for an element of the same language at all other levels (e.g., lexical selection may result in the choice of the element “potato”, in English, however at the phonological/phonemic level English aspirated /p/? may not reach selection and Spanish /p/? may be selected instead). This system would result in a coherent language choice across the linguistic system most of the times, yet it allows for the selection of elements of different languages across linguistic levels.

The question of how the selection device is implemented at each linguistic level is closely connected with how these same selection and switching devices are engaged: Do switches always occur top-down, such that individuals choose when the language is going to switch, or could it be that a combination of factors can trigger these switches contextually? The latter possibility would suggest that a switch process could be triggered contextually when a set of specifiable factors, albeit including stochastic variability, align (for sociolinguistic support of the latter see Auer, 1998; Woolard, 2004). The final principle that needs to be specified in a model of lexical access is the way in which output selection occurs. In essence, this could be achieved either strictly on the basis of the initial independent levels of activation that elements receive, or one may suggest that there is a need to invoke another mechanism, for example a suppression mechanism (Green, 1998b; Green & Abutalebi, 2013), to achieve selection. One implication of invoking an inhibition mechanism is that language production becomes inherently effortful, since at any given point inhibition is being applied to at least some subset of elements of the lexicon. If one proposes such a mechanism, then a question arises as to whether this principle governs all aspects of lexical selection: does the selection of any word rely on the suppression of the others? In other words, when I say chair, do I need to suppress ottoman, armchair and stool? If so, one could argue that inhibition is a general principle of how the mind and brain work, and consequently suppression in bilingualism would merely be a meta-tool extended to this particular case. However, if the claim were that this is a mechanism that applies specifically and only to bilingual language selection, then one would have to characterize its nature, how it comes about, the time line of its development, and the implications of such a particular tool for linguistic processing more generally. Further, and critically, it needs to be articulated how this inhibition principle unfolds over all levels of language: if at the lexical level all words in the non-target language are inhibited, does this mean that all syntactic frames that do not
belong to the target language are inhibited too? Are all the phonemes that are distinct across phonemic inventories suppressed? If, alternatively, inhibition at the lexical level is only applied to the direct competitor or translation equivalent and not to all lexemes in the non-target language, how does this principle generalize to linguistic levels where translation equivalents are rather unclear?

An alternative possibility for output selection that requires one fewer assumption is that the levels of activation of individual items are initially modulated such that they result in the maximum activity being received by the target element, bypassing the need to invoke an additional mechanism for successful lexical selection. The implication here would be that the parameter value or activation boost sent down from some Language node to any element that fulfills this criterion has to be large enough to make the other language not a real competitor when a speaker needs to commit to producing a single language. In such an architecture, the same operational principle could be applied to elements at all levels of language processing, sidestepping the implementation generalizability concerns of an additional device such as inhibition.

The combination of different assumptions for each of the discussed principles will result in theoretical models with divergent predictions for behavior. Here we propose one possible combination of such principles that we argue can capture bilingual behavior during natural communication to a greater extent than its predecessors.

3. Framework description

3.1. General selection mechanism

A critical question one needs to ask when devising any selection mechanism is how such a mechanism will work across linguistic levels. Intuitively, one would aim for a set of principles that can straightforwardly operate on all levels of representation. This is because otherwise one would have to explain how each mechanism developed exclusively for a particular level, and would need to characterize a different set of principles governing each of those levels. Despite this perhaps obvious observation, attested proposals of language selection in bilingual individuals have been mostly focused on the lexical level (Costa, Miozzo, & Caramazza, 1999; Green, 1998a; Green & Abutalebi, 2013; La Heij, 2005; for in-depth reviews and discussion of their problems see: Kroll & Gollan, 2014; Runnqvist et al., 2014). This has left the bilingual cognition literature lacking a proposal for a cohesive characterization of language selection across the whole linguistic system. In what follows, we describe a selection mechanism that can operate on the same principles across all linguistic levels, providing a unified account of language selection in bilingual individuals, and constituting the first characterization of the bilingual language system as a whole.

In this framework, the main principle that governs the selection of linguistic elements in bilingual individuals exclusively involves the selection of the most highly active item (whether this item is a sentential frame, lexical item, a morpheme or a phoneme). Crucially, then, the activity levels of candidates must be modulated such as to result in the highest activation level for the intended candidate, which is subsequently selected for production. We propose that the activation levels of all linguistic elements are determined by a combination of factors including: 1) frequency of each individual element in each language, 2) language proficiency of the speaker, 3) temporal effects (recency of use or decay in activation after use of both the item proper and of the individual features that constitute the item, including e.g., language (have I been speaking English up to this point) and register (e.g., have I been speaking in the polite voice; for instance, using the plural for a single interlocutor in Spanish (usted) or German (Sie), or have I been using the casual second person singular (tú in Spanish, du in German), 4) intended semantic meaning, and 5) communicative context. Importantly, these factors operate and determine activation levels for items at every linguistic level. We will zoom into each of the levels in Section 3.2 but to illustrate the functioning of the principles, we will use the lexical level as an example (see Fig. 1), since this is the level that previous proposals have attempted to describe and hence constitutes the easiest point of comparison.

The frequency of an element in each individual’s lexicon establishes the default activation levels of items (Fig. 1A), which include forms from both languages, and are modulated by the individual’s proficiency in each of the languages. This default distribution of activation is altered by temporal effects, which increase the activation levels of the most recently used features and forms (Fig. 1B). Semantic context extends activation to the nodes of both languages related to the intended message (Fig. 1C), whose weight is further modulated by the communicative context of the discourse (Fig. 1D). Note that even though the explanation of this process is sequential in the prose, we do not imply a sequential unfolding of activation, all these factors simultaneously spread activation to the language systems to modulate activation levels. Because of the spreading activation from the engaged features to all relevant elements, in cases in which the target form is not available or its level of activation does not reach selection threshold, the closest alternative candidates will be available for selection, including related elements in the same language and translation equivalents in the non-target language.

Communicative context includes, for example, higher availability of an element in a given language or finding a better match for the intended conceptual message in the form of one language over the other, as well as factors external to the speaker such as instructions to speak English or constraints imposed by the interlocutor and their language proficiency. The conceptualization of these factors’ influence on utterance selection is in many ways similar to the audience design considerations proposed by Ferreira (2019) for monolingual individuals; whereby known properties of the addressee (e.g., child versus adult status; here also language proficiency) or the message (e.g., emphasizing certain properties of the message over others) will determine word/structure selection and ultimately utterance production.

We postulate that the activation flow across linguistic levels is not channelled in a way such that only those elements belonging to the target language receive activation (La Heij, 2005; Macnamara, 1967). Instead, we propose that activation will flow freely to target and non-target languages, but that the nodes of the target language will receive additional activation from a Language feature, boosting their activation above those of the non-target language. This Language feature is conceptualized as a node at the semantic/conceptual level, which sends activation down in parallel and in a similar manner to other semantic, conceptual or contextual features; i.e., it will send activation down to all the elements that contain that feature (similar to the language feature described in Grainger & Dijkstra, 1992; Dijkstra & Van Heuven, 2002). This is to say that the language node English will spread activation down to the lexical elements “dog” and “cat” the same way that the semantic node Animal will spread activation down to those elements. In this way, the language system’s functional architecture in bilinguals is identical in all respects to that in monolinguals but for the simple addition of a Language node which functions like other properties of linguistic items such as whether the item belongs to a given register, dialect, etc., and it is represented via a direct link between each element at each linguistic level and a language node. In this framework, the activation levels of elements in the language system will result from the combination of the activation contributed by the following factors:

Baseline Frequency + Recency + Conceptual message + Communicative context/Language + Additional factors
In a communicative context in which the interlocutor only understands one language, the activation increase generated by the language boost will often effectively override the weight of the other factors, generally resulting in the selection of a linguistic form of the appropriate language. However, in a context in which the interlocutor understands both languages, the activation sent by both Language nodes will be equal. Thus, selection will be more heavily determined by other factors such as which element shares more features with the semantic level and hence expresses more accurately the target conceptual message, or how available or frequent the word is. Importantly, frequency here is assumed to be lower (in absolute terms) in bilinguals as compared to monolingual individuals, as suggested by the frequency-lag or weaker links hypothesis (Gollan et al., 2011; Gollan, Montoya, Cerna, & Sandoval, 2008; Gollan, Montoya, Fennema-Notestine, & Morris, 2005).

In short, this hypothesis holds that since bilinguals are exposed to and produce each language less frequently than monolinguals, the frequency of lexemes in both languages will be functionally lower, resulting in reduced or slower accessibility of lexemes both in their L2 relative to L1, but also in L1 as compared to monolingual (Gollan & Silverberg, 2003; Gollan, Montoya, & Werner, 2002; Gollan et al., 2005; Gollan et al., 2008; Sandoval, Gollan, Ferreira, & Salmon, 2010; Gollan et al., 2011; similar ideas in Ivanova & Costa, 2008; Lebomn & Laine, 2003; Mägiste, 1979; Nicodadis, Palmer, & Marentette, 2007; Ransdell & Fischler, 1987). Here we adopt this principle and extrapolate it to all other linguistic levels, including phonological, morphological and syntactic forms. Facilitation effects observed for shared forms across languages, such as cognate and homophone/homograph facilitation effects, would then straightforwardly follow from added cumulative frequency for such items over both languages.

Temporal effects such as recency and decay will of course also contribute to the levels of activation of all elements. This factor could account for difficulty in retrieving forms in one language after having used their equivalents in the other language for some period of time, since the activation levels of recently used terms will be boosted and the levels of the translations will have decayed over time.

3.2 Levels of representation and selection

3.2.1 Semantic level

The model proposed here distinguishes at least five levels of representation. The first level is the lexical-semantic network, which contains the properties that strictly constitute word meaning and several additional factors that combine to constrain lexical selection (Fig. 2, pink panel). The semantic features range from properties that are in the narrow sense fundamental to the meaning of a word (e.g., “furry”, “mammal”, “barks”), to broader conceptual/contextual features such as register, politeness and specialization. The boundaries between these different types of features are fuzzy: word register can be part of the core meaning of a word (e.g., being formal is a fundamental and distinctive feature of the meaning of the second person singular pronoun lei (formal) as compared to second person singular tu (informal) in Italian), but it does not have to be (i.e., the fact that a chair is standard register does not affect its denotation). For this reason, we do not establish a hard distinction between these different types of features and instead characterize a continuum from purely semantic to broader conceptual and contextual properties (Fig. 2, pink panel, left). These contextual features additionally include aspects such as semantic tempus and semantic number (i.e., are there one or two dogs, did the event happen today or yesterday), which do not constitute the core meaning of a word yet are part of the conceptual message and constrain and determine lexical and morphological selection. Finally, the language to be spoken by the interlocutor is also specified at this level, yet is outside of the realm of semantic properties (Fig. 2, pink panel, right). This node will become active when there is an active choice of language, which can be driven by contextual factors external to the speaker (e.g., a teacher asking the speaker to use English) or by internal factors such as the intention to place emphasis on a certain phrase or more faithfully replicate a third person’s speech (i.e., top-down switches). This contrasts with bottom-up switches, which are involuntary and emerge as a consequence of the combination of a number of factors such as availability, specificity, or stochastic variation internal to the speaker.

This proposed organization of semantic features assumes a shared semantic space between the different languages of multilingual individuals. Whether the meanings associated with particular words are shared across translation equivalents or whether each language has an independent storage has been a longstanding question in the bilingual literature (Pavlenko, 2009; Van Hell & De Groot, 1998). Here we propose that the semantic space is shared between languages (i.e., the semantic properties/features for each lexeme will be drawn from a shared pool), but we allow for semantic representations of translation equivalents to be associated with distinct features. Specifically, when the boundaries between concepts fully overlap across languages (e.g., for the concept dog) the selected features will be identical across languages. However, when semantic boundaries vary across languages, translation equivalent lexemes will have some overlapping and some distinct features, all drawn from the same semantic feature space.

For example, the lexeme cup will be associated with the semantic nodes “ceramic”, “has a handle”, “contains hot beverages”, “small”, “for drinking tea”, while its Spanish translation equivalent taza will pick some of the same features: “ceramic”, “has a handle”, “contains hot beverages”, but not others such as “small”, given that taza also englobes the meaning associated with mug in English.

3.2.2. Lexical level

Combined activation from these Semantic, Conceptual/Contextual and Language features cascades down (Caramazza, 1997; Dell, 1986; Morsella & Miozzo, 2002; Navarrete & Costa, 2005), in parallel and independently to modality specific phonological/orthographical lexemes (here construed as roots; see Halle & Marantz, 1993, 1994) and to grammatical properties from the morphosyntactic network. There are two consequences that follow from activation being cascading and independent. First, multiple units at the lexical and morphological level (e.g., those corresponding to the target and its neighbors) will send activation to the phoneme level, allowing the phonemes corresponding to the target’s related neighbors to also become active. Second, morphosyntactic properties will send subsequent activation down to the phonological level independent of lexical selection processes and vice versa, allowing for correct inflections and articles being retrieved even in the absence of successful lexical retrieval. Additionally, our model assumes a mostly non-interactive flow of activation, although we consider that if there were to be interactivity, such process would be reduced to the phonological level (as proposed by Rapp & Goldrick, 2000). Here, we will first characterize the lexeme level (Fig. 2, blue panel) and subsequently characterize the morphosyntactic network (Fig. 2, green panel).

In contrast to models suggesting the existence of an abstract, modality independent lemma as have other monolingual (Bock & Levelt, 1994; Dell, 1986; Levelt, Roelof, & Meyer, 1999) and bilingual (De Bot & Schreuder, 1993; Green, 1998b; Poulisse & Bongaerts, 1994) models, we propose modality specific (phonological/orthographical) lexemes (see Caramazza, 1997; Caramazza & Miozzo, 1997; Miozzo & Caramazza, 1997, 1998). This distinction is not immediately relevant for the theoretical claims developed here regarding bilingual lexical access and will not be considered further.1

---

1 The evidence for this claim comes from patient studies showing dissociations in semantic error patterns in oral production and reading (Caramazza & Hills, 1990), and oral production and writing (Rapp, Benzing, & Caramazza, 1997), against predictions from modality independent lemmas, which would anticipate parallel errors in both domains (for further issues and argumentation against modality independent lemmas see Caramazza, 1997; Caramazza & Miozzo, 1997; Miceli, Benvenuti, & Caramazza, 1997; Miozzo & Caramazza, 1997, 1998).
Fig. 2. Representation of the flow of activation between a fragment of the different levels of representation from semantic and language features to lexeme and morpho-syntactic networks and then on to phonological information. N = noun; V = verb; Adj = adjective; M = masculine; F = feminine; Pl = plural; Sng = singular; Dual = dual number.

An important issue that needs to be addressed at this point is whether there are direct connections between lexemes, or whether the conceptual level necessarily mediates the connections between them. This has become especially relevant in bilingual models, as researchers have suggested that access to the second language lexeme may occur through the first language translation equivalent. Here we posit that connections between translation equivalents occur primarily through the semantic level, particularly in highly proficient bilinguals. Concretely, in a context in which someone is translating a discourse unit larger than a single word, translation will occur via the conceptual system in the following manner: the comprehension of lexical items in one language will activate their corresponding semantic and conceptual/contextual nodes, which will in turn spread activation down to their translation equivalents. This route to translation should be uncontroversial at an intuitive level: successful translations from one language to another are unattainable on a word-by-word basis and rather rely on conceptual comprehension of the message that is successively formulated in the translated language. Interestingly, this connection through the semantic system also allows for the frequent use of unorthodox semantic calques that only make sense for bilingual individuals; for instance, “llamar para atrás,” (literally “to call backwards”), which is incorrect in Spanish but bilingual addressees understand as “to call back” through activation of the English meaning.

However, the model does not preclude the possibility that there are direct links between lexical items, too (in line with the Revised Hierarchical Model, Kroll & Stewart, 1994; Kroll, Van Hell, Tokowicz, & Green, 2010). Such a model may be motivated by the need to explain, for instance, that it is possible to learn that a new (non)word (e.g., gumpf) is the translation equivalent of another (e.g., bamp), without any true conceptual knowledge about the meaning of each lexeme. However, there is no operative reason for the existence of these links, and the implementation and consequences of these links are unclear at the current time. Furthermore, these associations could also be established through even extremely impoverished conceptual information associated with these words. For example, it could be that the (poor) conceptual representation of both gumpf and bamp is “the word that I learned from this list on this day”, and that those two words are related through this impoverished representation. Given that there is no direct evidence supporting one account over the other, we remain agnostic as to the existence of direct links between lexemes, but we argue that if they were to exist, there is no compelling reason to believe that these links would be in any way specific or special in bilingual individuals: presumably the exact same process and connections would be in place if one were told that gumpf and bamp are synonyms. Hence, we suggest that the relation between translation equivalent lexical items occurs through the conceptual level, but that if links were to be established between lexemes directly, the nature of these links would not fundamentally differ in the bilingual and in the monolingual lexicon.

3.2.3. Morphosyntactic level

The morphosyntactic network (Fig. 2, green panel) contains grammatical features that are organized in sub-networks such as word class (noun, verb, etc.), gender, and argument structure. The network contains as many sub-networks as exist in the combination of languages a multilingual understands. If an individual who only spoke English starts learning a gendered language such as French, a new sub-network for gender operations will be created within the shared morphosyntactic network. If an individual who spoke a language with only masculine/feminine gender distinctions learns a language that additionally has neutral gender, a new sub-network for gender operations will be created within the gender sub-network for this case. Each subnetwork or node within the subnetwork is connected to the node(s) of the language(s) that it can be realized in, and to the lexemes that hold that property (i.e., the lexeme sun will have a connection to the nodes noun and singular, while its translation equivalent sol will have a connection to the gender feature masculine in addition to connections to the same noun and singular nodes). The morphosyntactic network (and subnetworks within) are thus shared across languages. This proposal receives support from reports of bilingual patients whose lexical impairment affects only certain sub-networks (e.g., nouns) but, crucially, the impairment is qualitatively
equal across languages. For instance, an English and Arabic bilingual patient with acquired lexical access deficit showed an impairment in naming abstract as compared to concrete words in both languages (Crutch, Ridha, & Warrington, 2006); and in a study of Greek–English bilinguals with anomic aphasia, they were found to have comparable verb specific deficits in picture naming in both languages (Kambanaros & Van Steenbrugge, 2006). Importantly, the magnitude of the impairments may vary across languages (mostly showing benefits for the language patients are most fluent in; e.g., Kuzmina, Goral, Norvik, & Wekes, 2019). However, this quantitative difference does not bear a challenge for our model: in the same manner that low-frequency words are usually the most “damaged” in monolingual patients, they will also be the most unavailable in bilingual patients, it just so happens that for unbalanced bilinguals there is often a correlation between L2 proficiency and low-frequency elements.

Importantly, even though we suggest that the grammatical sub-networks and nodes are shared across languages, the morpho-phonological operations sanctioned by grammatical context are language-specific and take different forms for different languages (i.e., both languages share the grammatical feature of plurality, but it realizes as an “-s” for English and “-k” for Basque). This implies that when any given word undergoes a morpho-phonological transformation, the combination of activation flowing from the Language node and from the inherent properties of the lexeme (e.g., masculine gender) and from contextual semantic features (e.g., plural) will result in the appropriate transformation being applied. For example, an input to the morphophonological network of a combined activation from the nodes Spanish + gender feature (m) + contextual number (pl) will result in the output of the phonology /ɔ/. Although activation flows from semantic and conceptual/contextual features to grammatical operations, not all of the latter will receive activation from the semantic level: whether they do or do not will be determined by the inherent or contextual nature of the feature in question (Kibort & Corbett, 2008). Semantic features that have a corresponding grammatical expression will receive activation directly from the semantic or conceptual level (e.g., tense: did the message that I intend to express happen yesterday or today; number: more than one item), resulting in the corresponding morpho-phonological transformation to the relevant verb or noun; features that are lexically inherent (e.g., grammatical gender, plurality of trousers) will receive activation directly from lexical nodes. Last, purely intrinsic grammatical features will also be necessarily activated through lexical nodes, since they are an inherent property of the word itself (e.g., argument structure and inflectional class).

The independence of the cascading activation from lexical-semantic, contextual and conceptual features to the lexemes and to the morphosyntactic network is critical to account for two pervasive phenomena in speech production and bilingualism. First, it can explain the tip of the tongue phenomena where a person can retrieve some of the grammatical features of a target word (e.g., grammatical gender) while failing to produce its phonology (Caramazza & Miozzo, 1997; Miozzo & Caramazza, 1997; Vigliocco, Antonini, & Garrett, 1997), but also vice versa (Caramazza & Miozzo, 1997). Second, it can account for a ubiquitous phenomenon in fluent bilingual communication that has not been addressed by previous models of bilingual production: the application of morphophonological transformations of one language onto roots of the other language. For example, De Bot (1992) reported an instance where the French argument structure of the object, was used with the Dutch translation equivalent (Crutch, Ridha, & Warrington, 2006). Importantly, the magnitude of the impairments may vary across languages (mostly showing benefits for the language patients are most fluent in; e.g., Kuzmina, Goral, Norvik, & Wekes, 2019). However, this quantitative difference does not bear a challenge for our model: in the same manner that low-frequency words are usually the most “damaged” in monolingual patients, they will also be the most unavailable in bilingual patients, it just so happens that for unbalanced bilinguals there is often a correlation between L2 proficiency and low-frequency elements.

3.2.4. Phonological level
Once the appropriate lexeme and the relevant morphosyntactic operations have been computed, activation from the lexeme and the grammatical features combined cascades down to the phonological system, following the same activation flow principle as that from the semantic to the lexeme level (i.e., the amount of activation that spreads down will be proportionate to the corresponding lexeme’s activation; Caramazza, 1997; Cutting & Ferreira, 1999; Dell, 1986; Goldrick & Rapp, 2002; Griffin & Bock, 1998; Harley, 1993; Humphreys, Riddoch, & Quinlan, 1988; Rapp & Goldrick, 2000). At this level, the same selection principle is applied whereby the most active phonological form, as determined by the combination of the previously defined factors, will be selected for production. In the case of bilinguals, this will mean that in the majority of cases, the most active phonemes will be the adequate form as determined by the language of the selected lexeme. However, occasionally, the factors influencing activation levels (availability, context, etc.) will lead to the selection of some or all phonological elements of the other language. For instance, native speakers of Italian who are fluent in English will often fail to appropriately produce aspirated /h/ in English and will drop it instead, producing “have” as /av/ instead of /hav/. This stems from overall activation of this aspirated /h/ being very low in the shared phonological inventory, as this sound does not exist in Italian. Arguably, this is no different than the substitutions and different phonological realizations of the same lexeme that are observed in monolinguals as a consequence of context or register (e.g., metathesis of /spr/., as in wasp and grass, to /ps/ in African American Vernacular English; Thomas, 2007).

Importantly, these substitutions will only arise when two phonemes, one of each language, share some, and only some, similarities. If the very same phoneme is used in both languages, there will be no room for such substitutions, and if a phoneme is only used in one of the languages (e.g., Zulu clicks), these substitutions will be impossible (Best, McRoberts, & Sithole, 1988). Thus, very much like at the lexical level, in the absence of availability of the target candidate in the target language (either because an individual has not acquired that phoneme or because activation levels of that form do not reach selection threshold), bilinguals will produce the highest activated candidate, which will be the phoneme closest in phonological space, resulting sometimes in the production of phonemes that exist in the other language. For instance, if a Spanish-English bilingual is aiming to produce /p/ in potato, which is an aspirated voiceless bilabial stop, insufficient activation levels of this phoneme could lead to the production of /b/, the unaspirated voiceless bilabial stop, which is closest to the target phoneme and will arguably have the highest activation value as it is common to both English and Spanish, resulting in higher cumulative frequency (applying the same principle as at the lexical level for cognates, phonemes that are common to both languages will have higher overall frequency and consequently higher activation levels). It is worth mentioning that the knowledge and influence of a second language will on occasion lead to assimilation and dissimulation of phonetic categories within the shared phonological inventory (e.g., Caramazza, Yeni-Komshian, Zurif, & Carbone, 1973), often leading bilinguals to shift the pronunciation of a phoneme closer to the equivalent in the other language. This phenomenon is not unique to the phonological level - in fact it seems to be an overall property of bilingualism. It also exists, for instance, in lexico-semantic relations,
The syntactic level contains all the syntactic frames that are known to a person in the combination of languages a multilingual understands, hence making the syntactic system shared between languages (in concordance with Hartsuiker et al., 2004; c.f., De Bot (1992); Ullman, 2001). These frames have connections to the semantic nodes (Language, Register, Semantic content, etc.) in the same ways the elements in the rest of the levels do. Frames such as “subject + verb + prepositional phrase” for main clauses, which are shared across English and German (e.g., I went to the store), will have connections to both language nodes, but structures that are particular to one language will instead exclusively have a connection to their respective language node. For instance, causative phrases structured as “causal conjunction + subject + verb + object” (e.g., because I needed milk) will be connected to the English Language node, but the “causal conjunction + subject + object + verb” structure will instead be connected to the German Language node (e.g., weil ich Milch brauchte).

The selection of the syntactic frames will follow the same principles applied to other levels, that is, the most available frame will be selected, and frame activations will be determined by the combination of the relevant factors. Baseline activation will be determined by the frequency and prevalence of a structure in a language, which will then be modulated by recency effects, such that frames that have been recently used will increase in activation, becoming more likely to reach selection threshold. This will lead to speakers often choosing the structural frames that have just been used in conversation, regardless of whether the prior sentence was produced in the same language (e.g., Bock, 1986; Branigan, Pickering, & Celand, 1999; Pickering, Branigan, Cleland, & Stewart, 2000, for meta analysis see Mahowald, James, Futrell, & Gibson, 2016) or in a different language (Loebell & Bock, 2003; Meijer & Fox Tree, 2003; Schoonbaert, Hartsuiker, & Pickering, 2007). Syntactic frames will additionally receive activation from the communicative context or audience considerations (Ferreira, 2019) such that the complexity of the frame will match the difficulty projected to be understood by the listener. This is the same mechanism that is involved for monolinguals, whereby individuals choose harder structures for formal or academic settings and simpler structures when talking to babies (i.e., “baby talk”) or to foreign speakers who may potentially not understand the language well. Last, frames will also receive activation from the conceptual level, such that the chosen frame will best match the conceptual message (e.g., the active or passive voice depending on the message one is trying to transmit). As in other linguistic levels, the highest activated element may sometimes not be the adequate one for the intended language. For instance, English-French bilinguals have been reported to produce the “noun + adjective” French frame in English as opposed to the obligatory “adjective + noun” frame (Nicolaides, 2006), and Basque-Spanish bilinguals commonly place the adverbial conjunction “but” in the last position of the sentence as in Basque even when they are speaking Spanish, whose syntax requires it to be at the beginning of the clause (for further evidence see examples of syntactic transfer, e.g., Hohenstein, Eisenberg, & Naigles, 2006, Marian & Kauhansky, 2007).

Having developed the selection principle, one then has to address how the choice of frame subsequently constrains lexical, morphological and phonological activation. It follows from the description above that initially frames in both languages will be activated. If the frames overlap, it is expected that lexical and morphological elements in both languages will also be activated since they could potentially be inserted at any given point in the sentence. What happens though when the
boundaries of the syntactic units in the frames do not overlap? Will activation still spread across elements of the two languages?

We propose that this is in fact the case: initially activation will spread from both syntactic frames to their corresponding morphological, lexical and phonetic units, which enables situations such as the one described above where one can produce sentences in one language using the word order of the other. However, even though these cases are possible, elements from the non-target language are more likely to occur at those points of the syntactic frame where the boundaries overlap. Hence, we adopt a soft version of the equivalence constraint proposed by Poplack (1980), which predicts that language switches will be most likely at points where the surface structures of the languages coincide, or between sentence elements that are normally ordered in the same way by each individual grammar (for other proposals that predict more switches the more congruence there is between the two languages’ structures see: Deuchar, 2005; Moysken, 2006; Sebbas, 1998; Weinreich, 1953; for experimental work supporting this notion see Beatty-Martínez & Dussias, 2017; Her-ring, Deuchar, Parafita Couto, & Moro Quintanilla, 2010; Kootstra, Van Hell, & Dijkstra, 2010). However, even though the system prefers switches at shared boundaries, language switches seem to be allowed between constituents regardless of order, and within constituents at boundaries that do not exactly align but can be made to align by duplicating information. The example below illustrates both of these points. The sentence translates to “I told her that I wanted to come”. In Spanish, the natural word order first is main then relative clause, and these two are joined by the marker “que” at the beginning of the relative clause:

\[
\begin{align*}
\text{Sp:} & \quad \text{Le} \quad \text{he dicho} \quad \text{que} \quad \text{quería} \quad \text{venir.} \\
\text{Dat. (to her)} & \quad \text{(I) told} \quad \text{that} \quad \text{(I) wanted} \quad \text{to come}
\end{align*}
\]

In Basque, the syntactic order is reversed: first comes the relative clause, which is then followed by the main clause, and these two are joined by the marker “-la” attached to the auxiliary verb at the end of the relative clause.

\[
\begin{align*}
\text{Beq:} & \quad \text{Etorri nahi nuela} \quad \text{esan diot} \\
\text{Come wanted} & \quad \text{I-that} \quad \text{told} \quad \text{I to him/her}
\end{align*}
\]

Even though the order of the clauses is reversed, and that even within constituents the boundaries do not align (i.e., relative particle at the beginning of relative clause in Spanish but at the end in Basque), speakers will often produce sentences such as the following:

\[
\begin{align*}
\text{Mixed:} & \quad \text{Le} \quad \text{he dicho} \quad \text{que} \quad \text{etorri nahi nuela} \\
& \quad \text{I told him/her} \quad \text{that come wanted} \quad \text{I-that}
\end{align*}
\]

Where they use the Spanish syntactic frame Main clause + relative clause, and then within the relative clause, they include both the Spanish and the Basque markers, thus enabling the switch at a linguistically misaligned boundary. This suggests that both structures were being computed in parallel, and after having entered the relative clause in Spanish, when the switch to Basque happened, speakers are able to repair the “ungrammatical” switch by duplicating the relevant (morpho)syntactic information. This phenomenon has also been attested in corpus analyses, for instance, of English – Welsh bilinguals whereby bilingual speakers attached the Welsh verbal suffix -io to English verbs to address the Welsh requirement of markedness in verbalized nouns:

\[
\begin{align*}
\text{Mixed:} & \quad \text{I love-soaps} (I love soaps; Deuchar, 2005)
\end{align*}
\]

Importantly, and consistent with the rest of the proposal presented here, this phenomenon is not unique to a single linguistic level, but rather is a principle that replicates across all the language system. For instance, at the morphological level, this occurs in instances such as “weil ich getriggered wurde” (because I got triggered) for English and German bilinguals, where the middle verb is required to be morphologically marked as a participle, leading to the addition of the German participle affix -ge-, but the English verb “trigger” requires -ed to become a participle, leading to the use of both morphemes to satisfy all constraints. Lexically, switches also occur even when it leads to duplicated content; e.g., “the small manina” to mean “the small hand”, even though “manina” in Italian already has “small” as one of its attributes.

### 3.3. Language selection at each representational level

Intuitively, the most straightforward characteristic of the language selection process would perhaps have been one in which once a language has been selected at the lexical (or higher) level, that choice is kept through all the subsequent levels for the rest of the production process. However, the empirical reality, as already alluded, is such that lexemes from one language can be combined with morphemes from another language (e.g., los truckos), and then pronounced with the phonology of one of the two languages or even with mixed phonology (’Trakos/ with aspirated English /t/ but an open front unrounded vowel (Spanish /a/), instead of open-mid back unrounded vowel (English /ʌ/). Thus, as proposed here, it appears that language selection can be affected independently at each linguistic level (see Fig. 3). However, not all processes are equally prone to such midstream switching: In terms of representational levels, for instance, the syntactic system seems much more resistant to intrusions than the phonological level. The likely driver of these inequalities in permeability to language switches stems from the different extents of boundary alignment at different levels. Thus, we generalize the softer version of the equivalence constraint (Poplack, 1980) adopted for the syntactic level to all levels of language.

For smaller units (e.g., phonemes), boundaries overlap at every gap between two phonemes, hence enabling one-to-one substitutions and insertions so long as other constraints (e.g., similarity to target phoneme, Section 3.2.4) are met. At the lexical level, language switches are also relatively local, easily allowing for lexical insertions so long as the demands of the slot in which they will be inserted are met (e.g., conceptual equivalence across languages). However, as the units grow larger and the scope grows from local to distributed over items and time, the places at which switching is possible becomes narrower, with different elements establishing dependencies with each other and reducing the points at which switching could occur. The morphological level is a little less local and hence switching becomes somewhat less rampant, even though it is still possible as long as morphological constraints (e.g., being a noun or verb, animate or inanimate etc.) are met. If, however, some core morphological property (e.g., gender) exists in both languages but the parameters across languages do not align (for instance, a word is masculine in one language but feminine in the other), the switch at that boundary will not be possible, or at least very unlikely. For instance, even though code-switches at determiner-noun phrases are generally frequent (Dussias, 2001; Jake, Myers-Scotton, & Gross, 2005; MacSwan, 2005a, 2005b), a gender mismatch will make those switches extremely infrequent (e.g., flower in Spanish is feminine “la flor”, and in Italian is masculine “il fiore”; we suggest that a code switch such as “il flor” would be if not impossible, certainly extremely unlikely). Finally, the syntactic level, because it is the level with the largest boundaries and dependencies among items, is the least likely to sanction switches. Still, as discussed in Section 3.2.5, switches at the syntactic level still occur.

It should be noted that since language- or code-switching in natural conversation is the mere result of the combined activation received from different nodes, the expectation is that it is not cognitively or behaviorally costly. However, to the extent that it can be used as a communicative resource by multilingual environments when a target element is not available in the current language, there may be a cost associated with it (see Bultena, Dijkstra, & van Hell, 2015; and Fricke, Kroll, & Dussias, 2016 for slowed speech rate and cross-language phonological influence preceding code-switches). For instance, if an English-French
bilingual individual speaking English attempts to find the word for butterfly and this item does not reach selection threshold, they might produce the word papillon instead, which satisfies the semantic but not the language constraint required in the current context. This may lead to a delay in the production of papillon, or to a slower production of it, but critically the root of the cost will not be due to the language switch per se, but rather the language switch will be the consequence of the cost of retrieval. This delay arguably reflects the time from the point at which the speaker becomes aware that they are supposed to produce a word in English, and that papillon is not it, to the production of the closest yet non-target word. This suggests that subsequent to selecting the highest activated element, the system involves an assessment and monitoring of whether the retrieved phonological form satisfies contextual requirements, which we will develop in Section 3.4.

In sum, we argue that bilingual individuals have fully integrated linguistic systems across all linguistic levels. There is compelling evidence that the different linguistic levels of both languages are active in bilingual language use. Evidence of this simultaneous activation has been found at the lexical (Gullifer, Kroll, & Dussias, 2013, reviewed in Costa, 2005), morphosyntax (Hartsuiker et al., 2004; Hartsuiker & Pickering, 2008; Hartsdik, Branigan, & Pickering, 2011), and phonological levels (Hermans, Bongaerts, De Bot, & Schreuder, 1998; Hoshino & Kroll, 2008; Jared & Kroll, 2001; Midgley, Holcomb, Walter, & Grainger, 2008; Thierry & Wu, 2007), even when the phonological systems are completely distinct such as between a spoken and a sign language (Emmorey, Petrich, & Gollan, 2012; Van Hell, Ormel, Van der Loop, & Hermans, 2009; Morford, Wilkinson, Villwock, Pinar, & Kroll, 2011; for a review see Hanulová, Davidson, & Indefrey, 2011). This simultaneous activation persists even when only one language is at play (Colomé, 2001; Colomé & Miozzo, 2010; Costa, Caramazza, & Sebastian-Galles, 2006; Hermans et al., 1998; Pouliot, 1999), and when the interlocutor does not understand one of the languages (Casey & Emmorey, 2009; for reviews see Bialystok, Craik, & Gollan, 2009; Kroll, Bobb, & Wodniecka, 2006). The most parsimonious account for these effects is that there is no qualitative difference between items that belong to distinct languages over and above extant differences between different linguistic forms within a single language (e.g., register, dialect, baby talk). Thus, we propose that in the same manner that activation flows from one node to a related node within a language (e.g., Alario, Segui, & Ferrand, 2000; Caramazza, 1997; Costa, 2005; Dell, 1986; de Heij, Dirx, & Kramer, 1996; Levelt et al., 1999), activation also spreads across languages, and the selection process in bilingual language production unfolds on the same principles as it does during monolingual language production: it simply selects the most available candidate at each moment in time, without need for control mechanisms that work in combination with convenient tags or flags at convenient places. The assumptions described in this model not only account for the ubiquitous co-activation of the two languages of a bilingual individual at each linguistic level, but this co-activation, coordination, and influence of one language on the other is the natural prediction of such assumptions.

3.4. System-external executive control/verbal self-monitoring

As the reader will have noticed, this framework is one where the system does not have any built-in intelligence – once a certain input has been given, it will run through all the levels of the system, selecting the highest activated element at each level, until it reaches an output. However, it is possible that sometimes the reached output does not adjust to the environmental demands; hence, there ought to be a system in place to withhold such a response and potentially restart the search. This framework assumes that speakers can explicitly exert control at two points in the process: i) at the beginning of the process, such that based on information about the addressee/communicative situation speakers can top-down determine which specific features of meaning should be linguistically encoded, including what language/dialect/register these should be encoded in, and ii) at the output level, once the phonological form has been determined (see also Bock, 1986; Ferreira, 2019; Finkbeiner, Almeida, Janssen, & Caramazza, 2006; Miozzo & Caramazza, 2003). At an output level, it would permit the production of words only in the intended language as controlled by a general self-monitoring system, of the kind proposed to repair slips and to prevent the production of non-words (Dhoooge & Hartsuiker, 2010, 2012; Hartsuiker & Kolk, 2001). This executive control mechanism is thus external to the lexical selection system and is the same as is in place to accommodate any other idiosyncratic feature of mono or bilingual communicative situations (e.g., hold a swear word when not allowed in a context; producing child-directed speech; adjusting to experimental instructions etc.).

4. Contrasts to inhibitory models of bilingual language production

Our framework diverges from the (arguably) most influential inhibition based models in the field (originally Inhibitory Control Model (ICM), Green, 1986, 1998a, 1998b; subsequently developed in Abutalebi & Green, 2012; Green & Abutalebi, 2013), in two fundamental ways. First, our framework assumes that the monolingual and bilingual language systems operate under identical principles. Through the combination of parameters influencing activation levels and the simple selection mechanism exposed above, we have constructed a language architecture that will operate qualitatively similarly for any number of languages an individual may know. For this reason, constructing an utterance will be equally effortful/less for monolingual and bilingual individuals. This contrasts with Abutalebi and Green (2013) proposal, whereby they affirm that language production will always be more effortful for bilingual than for monolingual individuals.

“Selection follows activation of alternative possible candidates for expressing a message. In bilingual speakers, the demand to select an utterance despite ‘equifinality’ recurs in a repeated and sustained fashion. Accordingly, we infer that, in principle, language use in bilingual speakers increases the demand on the processes involved in utterance selection over and above those that are imposed on monolingual speakers” (Green & Abutalebi, 2013; Journal of Cognitive Psychology, 25:5, p. 516).

Second, our framework presents a mechanism that does not necessitate any additional operation for selection beyond the selection of the element with highest activation level. In contrast, Abutalebi and Green (2013), advocates for inhibition as the sine qua non for bilingual lexical access. However, it is currently unspecified where this mechanism is instantiated and how it operates. Specifically:

“The locus of suppression may be at the level of the language task schema itself or at the level of particular lexical or syntactic competitors. We also leave open the precise mechanism of suppression. It may be one that directly inhibits the competing representation. Alternatively, it may be one in which the target representation and competing representation are interconnected via mutual inhibitory links and so increasing the activation of the target leads to suppression of the competitor indirectly.” (Green & Abutalebi, 2013; Journal of Cognitive Psychology, 25:5, p. 519).

A useful and accurate account of bilingual language production requires a specified mechanism for selection with defined parameters, as opposed to a system that could be instantiated at any/all levels of selection and can suppress any/all elements in the other language’s lexicon, through either direct or indirect connections between elements. A subsidiary consequence of this lack of specification is that it is unclear how such a mechanism would generalize across linguistic levels to cover the whole language system, and how/when it would develop for bilingual individuals specifically.²

² Note that even if one assumes a competitive model like Roodofs (1992), where similar candidates compete for selection and higher competition results in harder selection, one would still be relying on the highest levels of activation for selection, and one would still not require suppression. This discussion has no bearing on the arguments about bilingual language organization developed here and will not be considered further here.
5. Disposing of the inhibition requirement

Even though, as pointed above, the actual implementation of a putative inhibition mechanism is rather unspecified, we believe that there is still value in discussing the empirical data that has been taken as support for this account. The working definition of inhibition that has been used in experimental work has been as follows. Each lemma is associated with a language tag (e.g., L1 or L2), and when a concept is activated, the lemmas (or lexemes) in both languages that are associated with that concept will become active. Since both translation equivalents have been activated, in order to achieve production in the target language, reactive (i.e., subsequent) inhibition will be applied to the non-target lexical nodes. Crucially, it is assumed that the greater the proficiency in a language, the stronger its activation will be and hence, the more strongly it will have to be inhibited in order to produce the target language when it is the less proficient one.

Support for this type of proposal has come mainly from language switching tasks, where bilingual individuals name stimuli either in the target lexical nodes. Crucially, it is assumed that the greater the proficiency in a language, the stronger its activation will be and hence, the more strongly it will have to be inhibited in order to produce the target language when it is the less proficient one.

Further, directly following the predictions of inhibition based accounts, it has been found that it takes longer to switch to the dominant L1 than to the non-dominant L2, presumably because overcoming the strong inhibition applied to the L1 is more effortful than overcoming the weak inhibition applied to the L2 (Jackson, Swainson, Cunnington, & Jackson, 2001; Meuter & Allport, 1999; Thomas & Allport, 2000). This was found not to be the case. Additionally, recent research has shown that switch-costs can additionally vary as a function of proficiency, predictability, and response preparation time (e.g., Costa & Santesteban, 2004; Verhoef et al., 2009) in ways that inhibition-based accounts do not predict. Costa and Santesteban found that switches were symmetrical for highly proficient bilinguals even when they switched between a dominant L1 and a weak L3, contrary to the assumptions of inhibition based models, which would predict that it should take longer to switch into the dominant L1 as a function of overcoming the strong inhibition applied to this language as compared to the weak inhibition applied to L3. Verhoef and colleagues also found that the occurrence of language switching symmetries or asymmetries was not determined by the relative language proficiency of the participants, but rather by the preparation time allowed between cue and stimulus presentation. Additionally, switching asymmetries can be created within a single language by asking participants to switch into either fast or slowly available words (i.e., high frequency or low frequency words; Finkbeiner et al., 2006). This obviously poses a problem for inhibition based accounts, since it would be hard to imagine how these fast and slowly available words could be tagged such as to categorically inhibit all the members of either of these groups.

In all then, although there is a cohesive body of evidence that seemingly provides support for inhibition based accounts, there is as consistent a body that stands directly in opposition to the predictions of this model (for detailed reviews of all evidence in favor and against both positions see Kroll & Gollan, 2014; Rumngvist et al., 2014). In what follows we present alternative accounts that can explain the results that have been taken to support the inhibition based accounts while keeping the selection mechanism proper as simple as possible.

6. Accounting for the observed effects

If switch-costs and switch-cost asymmetries can be created from stimuli that cannot be tagged (and hence categorically inhibited), and can be made to disappear by using univalent stimuli or by letting people alternate languages freely, the suppression hypothesis must be incorrect at some level: it predicts effects where none are found and effects are found where none are predicted. Where do these switch effects and switch cost asymmetries emerge from, though, if not suppression at the lexical level? In all likelihood from outside of the lexical system, given that the pattern of it being harder to switch into the dominant task is replicated across a whole range of tasks that hold no relation to language or lexica (e.g., Allport & Wylie, 2000; Campbell, 2005; Cherkasova, Manoach, Intriligator, & Barton, 2002; Ellefson, Shapiro, & Chater, 2006; Koch, Prinz, & Allport, 2005; Leboe, Whittlesea, & Miliken, 2005; Lemaire & Lacotteur, 2010). Although providing an account of these asymmetric effects of switching in different domains is beyond the scope of this paper, we will discuss a couple of options that would be compatible with our proposal for lexical selection and could account for extant evidence to a larger extent than the inhibition based accounts (see also Gilbert & Shallice, 2002 and Yeung & Monsell, 2003 for additional proposals).

One possibility is that outlined by Schneider and Anderson (2011), which suggests that the asymmetry arises from “impaired” performance after a difficult trial. In other words, the increased cost of returning to an easy task would emerge from the fact that the previous task was hard, and not from the act of returning per se. As predicted by this hypothesis, the authors found that an easy trial preceded by a difficult trial showed a delay in naming regardless of whether there was a task switch or not (Schneider & Anderson, 2011). This account applied to our case would successfully predict i) the asymmetries observed for language switching tasks in imbalanced bilinguals, since one can be considered the easy and the other the difficult naming task ii) the asymmetries observed when switching from low frequency to high frequency words, for the same

3 Additional support has been claimed to emerge from the so-called “long term inhibition” effect, whereby participants are delayed during naming in their L1 after having named items in their L2 but not the reverse (e.g., Linck, Kroll, & Sunderland, 2009; Misra, Guo, Bobb, & Kroll, 2012). However, how a mechanism for long term inhibition could be implemented, and what such a procedure would mean for general lexical access, has not been theoretically specified. Hence, in the absence of an explanation on how this effect explicitly supports the theory of inhibition, we will refrain from discussing it further.
Another explanation for the results may be the response selection account proposed by Caramazza and colleagues (Finkbeiner et al., 2006; Finkbeiner & Caramazza, 2006; Malon, Costa, Peterson, Vargas, & Caramazza, 2007; Miozzo & Caramazza, 2003). This account proposes that when a stimulus can afford two possible responses and tasks encourage participants to make these conscious decisions about response options, such as during language switching tasks, the speech production system makes both responses ready for the output system. Subsequently, one of the potential responses needs to be excluded based on the provided cues. When the cues are consistent across trials, the selection criterion is already established and the responses may be selected as soon as they become available. If naming-cues change, as they do in switch trials, some time may be necessary for participants to update the response selection criterion. Importantly, this proposal argues that if the response becomes available too quickly when there has been a shift in the response criterion, participants may temporarily block it before it is articulated to ensure that an error is not made. Hence, that response needs to be regenerated before it can be produced, counter-intuitively leading to a delay for responses that become available quickly (for a similar proposal see Balota, Law, & Zevin, 2000; for additional empirical support and further theoretical description and development see Dhooge & Hartsuiker, 2010, 2011, 2012).

Although this account is speculative, it succeeds at accounting for all the discussed phenomena. It explains asymmetries in bivalent stimuli, since switching into the dominant language may result in temporary blocking of the answer, because this answer would enter the buffer too quickly. It can also account for the lack of asymmetries when switching into univalent stimuli, since these stimuli afford a single response and no rejection would have to be made in the output buffer. This account would also predict the lack of asymmetries or switch costs in voluntary switch development see Dhooge & Hartsuiker, 2010, 2011, 2012).

In short, both Schneider and Anderson’s (2011) proposal and the response selection hypothesis account for phenomena beyond that explained by inhibition-based models. The extent to which either of these mechanisms is more generally applicable is outside the reach of this manuscript, but crucially they help undermine an inhibition model as a foundation for bilingual lexical selection and as a successful model for capturing the processes at play during bilingual communication.

7. Framework predictions

If all the principles that we have laid out were to accurately characterize the bilingual language production system, what are the behavioral effects we should expect to see? In a nutshell, we should expect parallel results for tasks that involve a single language or two languages, provided that we place the same communicative constraints on them. This means that all the landmark effects that have been identified as indexes of linguistic processing in monolingual individuals (N400, P600, frequency effects, etc.) should also replicate in bilingual individuals whether stimuli contain elements of a single language or of both languages.

For instance, if one were to build a connectionist model similar to those constructed by Dell and colleagues (e.g., Dell, Oppenheim, & Kittredge, 2008; Gordon & Dell, 2003), which included a language node and defined activation by the combination of factors described above, one would expect the outcome of the model to be sentences that contain code-switches that match the grammatical constraints proposed in 3.2.5.

At an experimental level, we should expect the same effects to be observed for translation equivalents in bilinguals and synonyms in monolinguals. The evidence is still scarce, but in a series of recent experiments Dylman and Barry (2018) convincingly found this to be the case during picture-word interference tasks. Through the course of 5 experiments, they showed remarkably similar facilitation effects when participants responded while presented with a synonym in the same language as a distractor word, as when the distractor word was a translation equivalent.

Additionally, we would expect similar results when participants are externally cued to switch between dialects and registers as when they are asked to switch between languages. Although more data is still required, initial data seems to support this hypothesis. First, Krik and colleagues (Kirk, Kempe, Scott-Brown, Philipp, & Declerck, 2018) tested two populations of German – Ocher Platt and English – Dundonian Cots bidialectal individuals in a dialect-switching task. In both cases, experimenters found switch costs analogous to those observed in traditional bilingual language switching tasks. Further, when they tested a monodialectal English group that they trained on Dundonian dialect, they found the same type of asymmetrical switch costs as has been found across languages of uneven proficiency. Further, Declerck and colleagues (Declerck, Ivanova, Grainger, & Dunabeitia, 2020) tested participants in both register-switching and language-switching tasks and they found that across participants, there was a positive correlation of overall register- and language-switch costs. Further, they found that the switch-costs for formal French, which was the language common across both tasks, were similar across the two switching tasks hence supporting the postulation that the language selection mechanism will operate in ways that are common to other types of within-language selection criteria.

4 The current versions of inhibition-based accounts make a clear categorical distinction between language selection, which is posited as a unique across-language process, and other types of within-language selection. Technically, one could develop inhibition hypotheses further to suggest that not only language, but also register and dialect operations are based on inhibitory accounts. However, once these different parameters combine and inhibition needs to be applied to intersections of them it becomes challenging to develop how such a mechanism would work. In other words, one can easily envision the extension from the current account to also include categorically suppressing the low register while activating the high register, in the same categorical manner as elements of language B are proposed to be inhibited when speaking in language A. However, would this imply that when one suppresses the low register, this applies to elements of both language A and language B, or would it instead be the case that language B has already been fully suppressed and then additional inhibition is applied to low register of language A? Would these multiple inhibitory mechanisms be nested within one another? If yes, what would the hierarchy of such mechanisms be? Alternatively, would the mechanism have 4 distinct subtypes/tags (i.e., Language A high register, Language A low register, Language B high register, Language B low register) and 3 of those would be inhibited every time one of them is to be activated? This issue grows exponentially the more aspects of language one wants to account for. Hence, although at surface level it seems like inhibition-based theories could also account for these register effects, it becomes apparent by trying to develop the mechanism by taking into account all the features that rule natural communication that it is actually far from trivial to extend inhibition to other features.
Further, given that we have placed the executive control outside the lexical system, and that it will monitor output to adapt to communication demands including task instruction, it is predicted that any guided lexical retrieval in which external constraints coerce output selection should show the same effects, regardless of whether these are within or across languages. This includes parallel effects for switching languages and switching between any two sets of instructions; such as between naming the color or the suit of a playing card (Blanco-Elorrieta & Pylkkänen, 2016a) or naming a picture (e.g., chair) vs the category of the picture (e.g., furniture; Deycler, Grainger, Koch, & Philipp, 2017).

A prediction that follows from arguing for integrated L1 and L2 language systems is that it should also be the case that bilingual individuals should have the easiest time speaking when allowed to use any item in their vocabulary, than when placing the constraint of having to stay in either language (hence effectively forbidding the use of half of it). In other words, we would expect to observe a benefit associated with enabling bilinguals to mix their languages. Very recent work also seems to point in this direction, showing that bilinguals are quickest in naming when allowed to mix languages at will (de Bruin et al., 2018; de Bruin et al., 2020).

At the syntactic level, one should expect that so long as ecological constraints are met while designing the stimuli and language switches occur at valid boundaries, syntactic violations within and across languages should elicit the same type of response (i.e., P600 effects). Similarly, we would expect that correctly formed sentences should not elicit such an effect even when the language of the sentence switches from one to the other; in other words, these sentences should be qualitatively processed as single-language sentences. Further, the expectation is also that the types of effects that have been identified above the single-word level in monolinguals, such as combinatorial processes during conceptual composition (e.g., LATL, Bemis & Pylkkänen, 2011; Blanco-Elorrieta & Pylkkänen, 2016b; Blanco-Elorrieta, Kastner, Emsmorey, & Pylkkänen, 2018b) should also replicate across languages in bilinguals.

Last, the combination of i) previous evidence that listeners constantly use available cues to predict and prepare for upcoming speech, and ii) the fact that in anticipation of switching languages bilinguals vary speech in a systematic way (i.e., they produce slowed speech rate and show cross-language phonological influence, Fricke et al., 2016), the prediction is that providing participants with these kinds of phonetic cues in the stimuli should allow participants to predict switch costs and reduce the processing load associated with them. Neuroimaging investigations of bilingual speech processing above the single-word level are still scarce, however, and finer-grained predictions will develop as this avenue of research provides more detailed characterizations.

8. Conclusions

The most important question in the bilingualism field has been how bilingual individuals manage to both communicate in one language without constant interference from the other, and freely switch between languages when the circumstances allow them to do so. Here we reviewed the principles under which any proposed bilingual language architecture could operate, and we present a framework of bilingual language organization that proposes common principles for element selection across all linguistic levels. This selection mechanism operates strictly on the basis of the highest levels of activation and does not assume an active suppression component. These activation levels are jointly determined by a conjunction of factors leading to highest activation of the target element, which is subsequently selected for production. A lexicon-external monitoring device then checks that the selected phonological form matches the criteria of the desired output. The proposed architecture describes phenomena occurring at every linguistic level and can account for attested features of bilingual speech both in and, crucially, also out of experimental settings.

Funding

This work was supported by the Dingwall foundation dissertation fellowship to E.B.E., and a Harvard MBB Provostial fund to A.C.

Declaration of competing interest

None.

Acknowledgements

This work is dedicated to the memory of Jacques Mehler – great scientist, gran signore, and beloved friend. I (AC) first met Jacques in the early 80s at a Sloan Workshop on language and aphasia in Nans-les-Pins. But I really got to know him well in 1986-1987 when together with an interesting cast of cognitive scientists (Walter Gerbino, Marc Jeannerod, Paolo Legrenzi, Pin Levelt, John Morton, Massimo Piattelli-Palmarini, Luigi Rizzi, Tim Shallice, and Paolo Viviani), we were involved in helping Daniele Amati, a theoretical physicist who had just accepted the directorship of the Scuola Internazionale Superiore di Studi Avanzati (SISSA, a type of grande ecole) in Trieste, develop a cognitive science/neuroscience program at the School. The group organized a series of meetings we called Trieste Encounters in Cognitive Science. Jacques was a polyglot – Spanish, French, Italian, English, and German – and he enjoyed speaking with his friends and colleagues in their language, if he could. During the more social parts of those meetings Jacques would switch from one language to another depending on his interlocuters.

I was to experience Jacques’ facile language switching far more intensely in 2001-2 when I spent a year at the SISSA where Jacques and Tim Shallice had taken full-time positions. The languages spoken on a daily basis were Italian, French Spanish and English, often involving many switches as when Jacques, Amati, Shallice, occasional visitors or students, and I would go for an aperitivo at one of the bars in Piazza Unità D’Italia, one street over from the then cognitive science laboratories. Jacques never seemed to tire when switching among his various languages. But if speaking one language involves having to suppress the others known by a speaker shouldn’t we expect that a poor polyglot, like Jacques, would have been perennially tired when speaking? He never seemed tired. Perhaps we should take that as “evidence” against those theories which assume that bilingual (polyglot) language production is intrinsically effortful due to the need to suppress competing languages. Jacques would have enjoyed this argument.

8. Conclusions

The most important question in the bilingualism field has been how bilingual individuals manage to both communicate in one language without constant interference from the other, and freely switch between languages when the circumstances allow them to do so. Here we reviewed the principles under which any proposed bilingual language architecture could operate, and we present a framework of bilingual language organization that proposes common principles for element selection across all linguistic levels. This selection mechanism operates strictly on the basis of the highest levels of activation and does not assume an active suppression component. These activation levels are jointly determined by a conjunction of factors leading to highest activation of the target element, which is subsequently selected for production. A lexicon-external monitoring device then checks that the selected phonological form matches the criteria of the desired output. The proposed architecture describes phenomena occurring at every linguistic level and can account for attested features of bilingual speech both in and, crucially, also out of experimental settings.


