ARTICLE IN PRESS

Cognition xxx (xxxx) xxx

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

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

Esti Blanco-Elorrieta[a,*], Alfonso Caramazza[b]

[a] Department of Psychology, Harvard University, Cambridge, MA, USA
[b] Center for Mind/Brain Sciences (CIMeC), University of Trento, Trento, Italy

ARTICLE INFO

Keywords:
Lexical access
Language selection
Selection-by-activation
Bilingual language organization
Bilingual communication
Bilingual language production

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 cohabitate 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 (e.g., Bloem & La Heij, 2003; Bloem, van den Boogaard, & La Heij, 2004; Costa, 2005; Green, 1998a, 1998b; Green & Abutalebi, 2013; Kroll & Gollan, 2014; La Heij, 2005), with some work on the syntactic (e.g., Hartsuiker, Pickering, & Veltkamp, 2004; MacWhinney, 2005) and phonological levels (e.g., Sebastián-Gallés & Bosch, 2005; Sebastián-Gallés & Kroll, 2003). 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

☆ This paper is a part of special issue "Special Issue in Honour of Jacques Mehler, Cognition's founding editor".

* Corresponding author.
E-mail addresses: blancoelorrieta@fas.harvard.edu (E. Blanco-Elorrieta), caram@wjh.harvard.edu (A. Caramazza).

https://doi.org/10.1016/j.cognition.2021.104625
Received 9 July 2020; Received in revised form 8 January 2021; Accepted 5 February 2021

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 this 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 to be activated would be one where there is a discrete switch that works as a floodgate by letting activation flow exclusively to the words that meet the language criterion. La Heij and colleagues, for instance, proposed this type of architecture (2005; Bloem & La Heij, 2003; Bloem et al., 2004), which would automatically result in selection from the appropriate language pool. However, experimental evidence seems to favor alternative accounts were words in both languages receive activation even when only one is required for communication (e.g., Martin, Dering, Thomas, & Thierry, 2009). Proponents of this latter type of activation flow are, for example, Costa (2005), Green, 1998a, 1998b; Green & Abutalebi, 2013), and Kroll and Gollan (2014); for a thorough review see: Runnqvist, Strijkers, & Costa, 2014), although it is unspecified what the mechanism could be that allows for activation to flow to elements of both languages. One possibility for such a system would be one where there is a switch of some kind (e.g., a Language node) at the highest level of the architecture that modulates activation levels by, for example, sending an activation boost to the words that meet the language criteria. Importantly, this is not an "all or none" system: this structure still allows for words in the non-target language to receive activation from semantic or conceptual nodes.

The shared mechanistic assumption in these accounts though appears to be that if a language switch occurs at one level of language production (e.g., the lexical level), this language choice will automatically coerce language choice in the rest of the linguistic levels (i.e, morphosyntax, phonology), through some unspecified operation, making language choice univalent for the whole linguistic system. Alternatively, items at each linguistic level could have their own connections to the Language node/switch, individually receiving activation when a decision is made as to what element to produce. This would mean that the activation boost sent by the Language node would independently reach elements at each linguistic level, and while it could boost the activation of an element of one language to selection threshold at one level (e.g., the lexical level), it would not necessarily boost activation to selection threshold for an element of that 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/phonetic level English aspirated /p^h/ 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

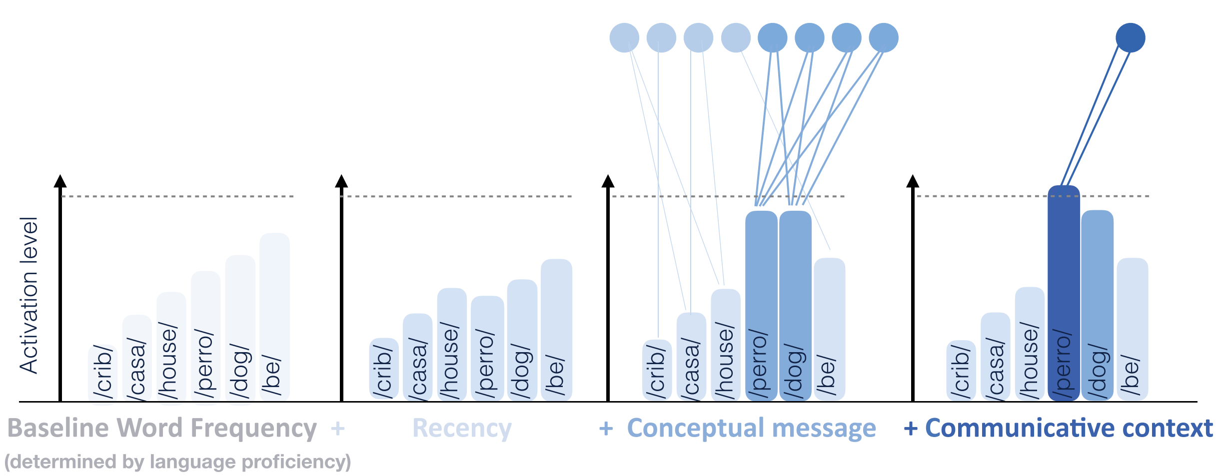

Fig. 1. Example of the selection principles applied to the lexical selection process. Node activation at any given point in time is simultaneously determined by a combination of (at least), baseline frequency, recency of use, overlap with the conceptual message to be transmitted and the communicative context of the moment.

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 system 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 channeled 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, Cera, & 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 monolinguals (Gollan & Silverberg, 2001; 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; Lehtonen & Laine, 2003; Mägiste, 1979; Nicoladis, 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, Roelofs, & Meyer, 1999) and bilingual (De Bot & Schreuder, 1993; Green, 1998a; Poulisse & Bongaerts, 1994) models, we propose modality specific (phonological/orthographic) 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 & Hillis, 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, Benvegnù, Capasso, & Caramazza, 1997; Miozzo & Caramazza, 1997, 1998).

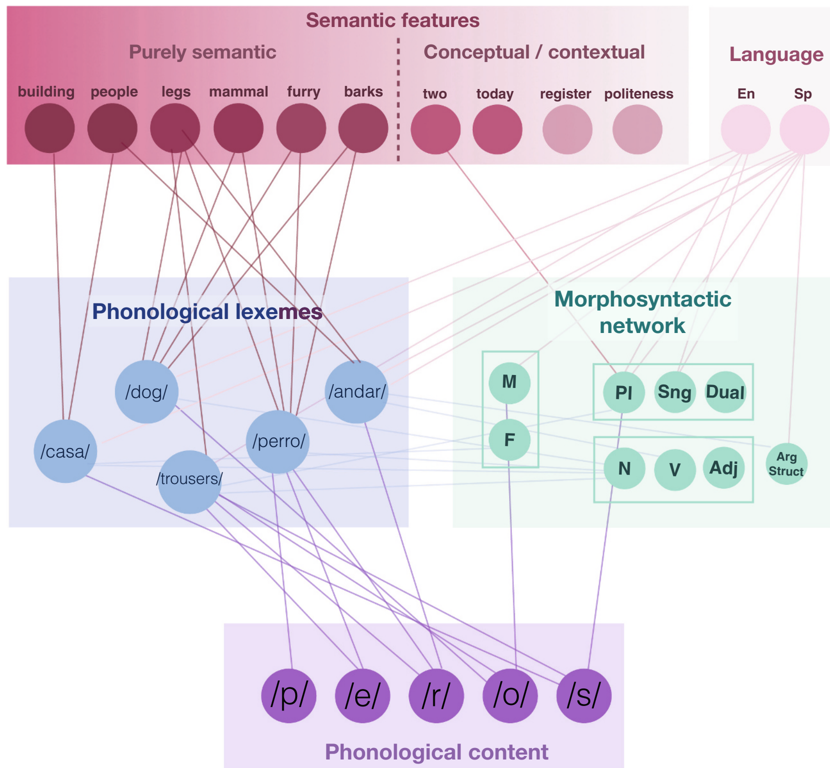

Fig. 2. Representation of the flow of activation between a fragment of the different levels of representation from semantic and language features to lexeme and morpho-syntactic networks and then on to phonological information. N = noun; V = verb; Adj = adjective; M = masculine; F = feminine; Pl = plural; 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 node 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, & Weekes, 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 /os/.

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 *voter* ("to vote"), which takes a noun phrase object, was used with the Dutch translation equivalent *stemmen*, which takes a prepositional object. Pfaff (1979) reports frequent morphophonological transformations of English roots with Spanish grammatical features; including verbs (e.g., "*ya no le* train*eará*", derived from English *to train* + Spanish verbal suffix "*-ear*" + Spanish future marking "*á*", meaning "s(he) will not train him/her anymore"), adjectives ("muy *conservativa*", derived from English adjective *conservative* + Spanish feminine gender marking "*a*"), and nouns (e.g. *los* truck*os*, derived from English *truck* + Spanish masculine gender marking "*o*" + Spanish plural marking "*s*"). This last example is particularly relevant because it illustrates 1) the independence of the cascading activation to phonological lexemes and grammatical features, as it shows that the person who produced this noun phrase could retrieve the gender feature associated with the target Spanish lexeme *camion* (masculine gender, meaning = *truck*), even if the actual phonology of the lexeme *camion* did not become available; and, 2) the extent to which the morphosyntactic network is shared between languages, since morphophonological transformations can be applied to roots of either language so long as they fulfill the relevant word class and phonological requirements (i.e., plural features require singular nouns to be applied on, but these operations may be applied on singular nouns of either language; see also Fig. 3).

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 /æv/ instead of / hæv/. 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 /sp./, as in *wasp* and *grasp*, 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^h/ in *potato,* which is an aspirated voiceless bilabial stop*,* insufficient activation levels of this phoneme could lead to the production of /p/, 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 dissimilation of phonetic categories within the shared phonological inventory (e.g., Caramazza, Yeni-Komshian, Zurif, & Carbone, 1973), often leading bilinguals to shift the pronounciation 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,

- Berak gauza piloa dauzka esku artean, que si uni, que si volunteering
(Bsq: (s)he has a lot of things in her hands, Sp: if uni, if Eng: volunteering)

- A si?
(Sp: really?)

- Bai, danatarik
(Bsq: Yes, all sorts of stuff)

- Eta apply egingo du, programme-era?
(Bsq: And Eng: apply Bsq: going to, Eng: programme Bsq: to the?)

- Ez, ez dut pentsatzen. Si acaso gurasoak asko insistituz gero baina ez dut uste.
Bsq: No, no I don't think so. Sp: If anything Bsq: if parents insist a lot but I don't think so.

- Nahi du lan egin social issue-tan eta horrelakoetan, si la
Bsq: He wants to work Eng: social issue Bsq:-in and in that sort of stuff, Sp: if the

policia harass-ea a los homeless eta horrela.
police Eng: harass Sp: -es the Eng: homeless Bsq: and such.

- Ah ya, orduan igual activism-ean egotea hobe grad school-en egotea baino.
Bsq: Ah I see, then maybe Eng: activism Bsq: -in to be better Eng: grad school Bsq: -in to be than.

Fig. 3. Example of fluent lexical, grammatical and phonological code-switching in Basque (blue) – Spanish (red) - English (green) trilinguals taken from a conversation between two informants.

whereby concepts in one language acquire properties of the same concept in the other language, which results in conceptual "blends" (e.g., Ameel, Malt, Storms, & Van Assche, 2009). However, this fact does not influence the tenets of the architecture proposed here. Even if the VOT boundary for stop consonants in English shifted as a consequence of also speaking French in English-French bilinguals, and these bilinguals produced /b/s in each language that are more similar to each other than the /b/s an English monolingual and a French monolingual would produce, these individuals still have an individual representation of each phoneme (perhaps even as allophones of the same phoneme), and the point still stands that if one of these forms is more frequent across languages, it will end up being chosen even when there is a less-frequent yet more-accurate phonological candidate in the target language.

3.2.5. Syntactic level

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 sematic 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, & Cleland, 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 (Nicoladis, 2006), and Basque-Spanish bilinguals commonly place the adversative 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 & Kaushanskaya, 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; Muysken, 2000; Sebba, 1998; Weinreich, 1953; for experimental work supporting this notion see Beatty-Martínez & Dussias, 2017; Herring, 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 is first main then relative clause, and these two are joined by the marker "que" at the beginning of the relative clause:

Sp:	Le	he dicho	|	**que**	queria	venir.
	Dat. (to her)	(I) told	|	that	(I) wanted	to come
	main clause		|		***relative clause***	

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.

Bsq:	Etorri	nahi	nue**la**	|	esan	diot
	Come	wanted	I-that	|	told	I to him/her
	relative clause			|	***main clause***	

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:

Mixed:	Le	he	dicho	|	**que**	etorri	nahi	nue**la**
	I	told	him/her	|	**that**	come	wanted	I-**that**
	main clause			|	***relative clause***			

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:

Mixed: dw i'n love-io soaps (I love soaps; Deuchar, 2005).

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 **ge**triggere**d** 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* truck*os*), and then pronounced with the phonology of either one of the two languages or even with mixed phonology (/t^{h}rakos/ with aspirated English /t/ but an open front unrounded vowel (Spanish /a/), instead of open-mid back unrounded vowel (English /^/). 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 in 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), morphosyntactic (Hartsuiker et al., 2004; Hartsuiker & Pickering, 2008; Hatzidaki, 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, Piñar, & 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, 2000; Hermans et al., 1998; Poulisse, 1999), and when the interlocutor does not understand one of the languages (Casey & Emmorey, 2009; for reviews see Bialystok, Craik, Green, & 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; La Heij, Dirkx, & Kramer, 1990; 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 more 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 (Dhooge & 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 back 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, 2013; 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 Roelofs (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 same or in a different language than in the previous trial.[3] Participants are slower on switch trials, and this delay is argued to reflect the time it takes to overcome the inhibition that was applied to the now target language on the previous trial (Costa & Santesteban, 2004; Costa, Santesteban, & Ivanova, 2006; Meuter & Allport, 1999; Thomas & Allport, 2000). 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; Philipp, Gade, & Koch, 2007; Schwieter & Sunderman, 2008; Verhoef, Roelofs, & Chwilla, 2009). These results have been widely replicated in multilingual individuals with diverse linguistic backgrounds (Blanco-Elorrieta & Pylkkänen, 2016a; Calabria, Branzi, Marne, Hernández, & Costa, 2015; Calabria, Hernández, Branzi, & Costa, 2012; Costa & Santesteban, 2004; De Baene, Duyck, Brass, & Carreiras, 2015; Declerck, Koch, & Philipp, 2012; Kang et al., 2017; Macnamara, Krauthammer, & Bolgar, 1968; Meuter & Allport, 1999), providing abundant data consistent with the inhibition based accounts.

However, although this evidence on the surface seems to make a compelling case in favor of an inhibition-based account of lexical selection, there are a number of factors that stand directly against it, and the biggest challenge comes from the very same switching paradigm that provides the strongest evidence for it. To start with, switch-costs are *only* obtained for bivalent stimuli (i.e., stimuli that are named in two languages during the experiment) and disappear when stimuli are assigned exclusively to one naming language in the course of the experiment (Finkbeiner et al., 2006; Gollan & Ferreira, 2009). In other words, switching into, e.g., L1, is costless if the stimulus type (i.e., picture, dot pattern) in question is always named in L1. Since inhibition-based accounts expect costs to arise from overcoming inhibition at the lemma level, the prediction would be for there to always be a switch-cost when one switches from one language to another, regardless of whether the target element is bi- or univalent. However, this was found not to be the case. Additionally, recent research has shown that switch-costs decrease with the use of more naturalistic cues (Blanco Elorrieta & Pylkkänen, 2015; Blanco-Elorrieta & Pylkkänen, 2017). It is unclear why using more naturalistic associations between cues and targets should result in a decrease or disappearance of the switch-costs, if switch-costs emerged because of inhibition at the lemma level. Along the same line, switch-costs and asymmetries also disappear when language switching is voluntary (i.e., when the participant can freely decide what language to use; Blanco-Elorrieta & Pylkkänen, 2017; Gollan & Ferreira, 2009; Kleinman & Gollan, 2016).

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; Runnqvist 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, & Milliken, 2005; Lemaire & Lecacheur, 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, & Sunderman, 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.

reason and iii) the lack of asymmetries in balanced bilinguals, since retrieval in both languages should be of equal difficulty. Although this account does not make specific claims about the effects of switches being voluntary, it could be argued that in this type of task all trials will be "easy", since participants will always choose to name in the easiest language, leading to no asymmetries or costs when switching. However, it is hard to think of an argument within this framework to account for the finding that asymmetries disappear when switching into univalent stimuli. Given that under this hypothesis delays are caused by the difficulty of the previous task, whether participants switch into bivalent versus univalent stimuli should not differentially affect reaction times.

Another explanation for the results may be the response selection account proposed by Caramazza and colleagues (Finkbeiner et al., 2006; Finkbeiner & Caramazza, 2006; Mahon, 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 these 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 switching, since there is no selection/rejection criterion to follow and the first response to enter the output buffer will at all moments be adequate to be produced. It could even account for recent findings showing that switching a language "on" (switching from producing a single language to producing two) is effortless but switching a language "off" (switching from double to single language production) is not (Blanco-Elorrieta, Emmorey, & Pylkkänen, 2018a). Since both responses are allowed in the switching "on" condition no rejection needs to happen following the activation of both possibilities, but a new rejection criterion needs to be applied when going from the simultaneous production of two languages to single language production, causing there to be a cost in this scenario. Last, since this account is concerned with response selection, and not really with language per se, it could easily be extrapolated to the behavioral tasks outside language that have shown similar patterns of results.

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 accounts is ultimately accurate 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 – Öcher 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, & Duñabeitia, 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]

[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; Declerck, 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, Emmorey, & 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 (ie., 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, Pim 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 école) 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.

References

Abutalebi, J., & Green, D. (2013). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics, 20*(3), 242–275.

Alario, F. X., Segui, J., & Ferrand, L. (2000). Semantic and associative priming in picture naming. *The Quarterly Journal of Experimental Psychology: Section A, 53*(3), 741–764.

Allport, A., & Wylie, G. (2000). Task switching, stimulus-response bindings, and negative priming. *Control of Cognitive Processes: Attention and Performance XVIII*, 35–70.

Ameel, E., Malt, B. C., Storms, G., & Van Assche, F. (2009). Semantic convergence in the bilingual lexicon. *Journal of Memory and Language, 60*(2), 270–290.

Auer, P. (1998). *Code-switching in conversation*. London: Routledge.

Balota, D. A., Law, M. B., & Zevin, J. D. (2000). The attentional control of lexical processing pathways: Reversing the word frequency effect. *Memory & Cognition, 28* (7), 1081–1089.

Beatty-Martínez, A. L., & Dussias, P. E. (2017). Bilingual experience shapes language processing: Evidence from codeswitching. *Journal of Memory and Language, 95*, 173–189.

Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience, 31*(8), 2801–2814.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14*(3), 345.

Bialystok, E., Craik, F. I., Green, D. W., & Gollan, T. H. (2009). Bilingual minds. *Psychological Science in the Public Interest, 10*(3), 89–129.

Blanco Elorrieta, E., & Pylkkänen, L. (2015). Brain bases of language selection: MEG evidence from Arabic-English bilingual language production. *Frontiers in Human Neuroscience, 9*, 27.

Blanco-Elorrieta, E., & Pylkkänen, L. (2016a). Bilingual language control in perception versus action: MEG reveals comprehension control mechanisms in anterior cingulate cortex and domain-general control of production in dorsolateral prefrontal cortex. *Journal of Neuroscience, 36*(2), 290–301.

Blanco-Elorrieta, E., & Pylkkänen, L. (2016b). Composition of complex numbers: Delineating the computational role of the left anterior temporal lobe. *NeuroImage, 124*, 194–203.

Blanco-Elorrieta, E., & Pylkkänen, L. (2017). Bilingual language switching in the lab vs. in the wild: The spatio-temporal dynamics of adaptive language control. *The Journal of Neuroscience, 37*(37), 9022–9036.

Blanco-Elorrieta, E., Emmorey, K., & Pylkkänen, L. (2018a). Language switching decomposed through MEG and evidence from bimodal bilinguals. *Proceedings of the National Academy of Sciences, 115*(39), 9708–9713.

Blanco-Elorrieta, E., Kastner, I., Emmorey, K., & Pylkkänen, L. (2018b). Shared neural correlates for building phrases in signed and spoken language. *Scientific Reports, 8* (1), 1–10.

Bloem, I., & La Heij, W. (2003). Semantic facilitation and semantic interference in word translation: Implications for models of lexical access in language production. *Journal of Memory and Language, 48*(3), 468–488.

Bloem, I., van den Boogaard, S., & La Heij, W. (2004). Semantic facilitation and semantic interference in language production: Further evidence for the conceptual selection model of lexical access. *Journal of Memory and Language, 51*(2), 307–323.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18* (3), 355–387.

Bock, K., & Levelt, W. J. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984).

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (1999). Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin & Review, 6*(4), 635–640.

Bultena, S., Dijkstra, T., & van Hell, J. G. (2015). Switch cost modulations in bilingual sentence processing: Evidence from shadowing. *Language, Cognition and Neuroscience, 30*(5), 586–605.

Calabria, M., Hernández, M., Branzi, F. M., & Costa, A. (2012). Qualitative differences between bilingual language control and executive control: Evidence from task-switching. *Frontiers in Psychology, 2*, 399.

Calabria, M., Branzi, F. M., Marne, P., Hernández, M., & Costa, A. (2015). Age-related effects over bilingual language control and executive control. *Bilingualism: Language and Cognition, 18*(1), 65–78.

Campbell, J. I. (2005). Asymmetrical language switching costs in Chinese–English bilinguals' number naming and simple arithmetic. *Bilingualism: Language and Cognition, 8*(1), 85–91.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology, 14*(1), 177–208.

Caramazza, A., & Hillis, A. E. (1990). *Where do semantic errors come from?Cortex,26, 95–122*.

Caramazza, A., & Miozzo, M. (1997). The relation between syntactic and phonological knowledge in lexical access: Evidence from thetip-of-the-tongue'phenomenon. *Cognition, 64*(3), 309–343.

Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., & Carbone, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *The Journal of the Acoustical Society of America, 54*(2), 421–428.

Casey, S., & Emmorey, K. (2009). Co-speech gesture in bimodal bilinguals. *Language & Cognitive Processes, 24*(2), 290–312.

Cherkasova, M. V., Manoach, D. S., Intriligator, J. M., & Barton, J. J. (2002). Antisaccades and task-switching: Interactions in controlled processing. *Experimental Brain Research, 144*(4), 528–537.

Colomé, A. (2001). Lexical activation in bilinguals' speech production: Language-specific or language-independent? *Journal of Memory and Language, 45*, 721–736.

Colomé, À., & Miozzo, M. (2010). Which words are activated during bilingual word production? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36* (1), 96.

Costa, A. (2005). Lexical access in bilingual production. In J. F. Kroll, & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 308–325). New York: Oxford University Press.

Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language, 50*(4), 491–511.

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection? *Journal of Memory and Language, 41*, 365–397.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1283.

Costa, A., Santesteban, M., & Ivanova, I. (2006). How do highly proficient bilinguals control their lexicalization process? Inhibitory and language-specific selection mechanisms are both functional. *Journal of Experimental Psychology-Learning Memory and Cognition, 32*(5), 1057–1074.

Crutch, S. J., Ridha, B. H., & Warrington, E. K. (2006). The different frameworks underlying abstract and concrete knowledge: Evidence from a bilingual patient with a semantic refractory access dysphasia. *Neurocase, 12*(3), 151–163.

Cutting, J. C., & Ferreira, V. S. (1999). Semantic and phonological information flow in the production lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(2), 318.

De Baene, W., Duyck, W., Brass, M., & Carreiras, M. (2015). Brain circuit for cognitive control is shared by task and language switching. *Journal of Cognitive Neuroscience, 27*(9), 1752–1765.

De Bot, K. (1992). A bilingual production model: Levelt's' speaking model adapted. *Applied Linguistics, 13*(1), 1–24.

De Bot, K., & Schreuder, R. (1993). Word production and the bilingual lexicon. *The Bilingual Lexicon, 191*, 214.

de Bruin, A., Samuel, A. G., & Duñabeitia, J. A. (2018). Voluntary language switching: When and why do bilinguals switch between their languages? *Journal of Memory and Language, 103*, 28–43.

de Bruin, A. M. T., Samuel, A., & Duñabeitia, J. A. (2020). Examining bilingual language switching across the lifespan in cued and voluntary switching contexts. *Journal of Experimental Psychology: Human Perception and Performance, 46*(8), 759–788.

Declerck, M., Koch, I., & Philipp, A. M. (2012). Digits vs. pictures: The influence of stimulus type on language switching. *Bilingualism: Language and Cognition, 15*(4), 896–904.

Declerck, M., Grainger, J., Koch, I., & Philipp, A. M. (2017). Is language control just a form of executive control? Evidence for overlapping processes in language switching and task switching. *Journal of Memory and Language, 95*, 138–145.

Declerck, M., Ivanova, I., Grainger, J., & Duñabeitia, J. A. (2020). Are similar control processes implemented during single and dual language production? Evidence from switching between speech registers and languages. *Bilingualism: Language and Cognition, 23*(3), 694–701.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*(3), 283.

Dell, G. S., Oppenheim, G. M., & Kittredge, A. K. (2008). Saying the right word at the right time: Syntagmatic and paradigmatic interference in sentence production. *Language & Cognitive Processes, 23*(4), 583–608.

Deuchar, M. (2005). Congruence and Welsh-English code-switching. *Bilingualism, 8*(3), 255.

Dhooge, E., & Hartsuiker, R. J. (2010). The distractor frequency effect in picture–word interference: Evidence for response exclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(4), 878.

Dhooge, E., & Hartsuiker, R. J. (2011). The distractor frequency effect in a delayed picture-word interference task: Further evidence for a late locus of distractor exclusion. *Psychonomic Bulletin & Review, 18*(1), 116–122.

Dhooge, E., & Hartsuiker, R. J. (2012). Lexical selection and verbal self-monitoring: Effects of lexicality, context, and time pressure in picture-word interference. *Journal of Memory and Language, 66*(1), 163–176.

Dijkstra, A. F. J., & Van Heuven, W. J. (2002). *The architecture of the bilingual word recognition system: From identification to decision*.

Dussias, P. E. (2001). Psycholinguistic complexity in codeswitching. *International Journal of Bilingualism, 5*(1), 87–100.

Dylman, A. S., & Barry, C. (2018). When having two names facilitates lexical selection: Similar results in the picture-word task from translation distractors in bilinguals and synonym distractors in monolinguals. *Cognition, 171*, 151–171.

Ellefson, M. R., Shapiro, L. R., & Chater, N. (2006). Asymmetrical switch costs in children. *Cognitive Development, 21*(2), 108–130.

Emmorey, K., Petrich, J. A., & Gollan, T. H. (2012). Bilingual processing of ASL–English code-blends: The consequences of accessing two lexical representations simultaneously. *Journal of Memory and Language, 67*(1), 199–210.

Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annual Review of Psychology, 70*, 29–51.

Finkbeiner, M., & Caramazza, A. (2006). Now you see it, now you don't: On turning semantic interference into facilitation in a Stroop-like task. *Cortex, 42*(6), 790–796.

Finkbeiner, M., Almeida, J., Janssen, N., & Caramazza, A. (2006). Lexical selection in bilingual speech production does not involve language suppression. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1075.

Fricke, M., Kroll, J. F., & Dussias, P. E. (2016). Phonetic variation in bilingual speech: A lens for studying the production–comprehension link. *Journal of Memory and Language, 89*, 110–137.

Gilbert, S. J., & Shallice, T. (2002). Task switching: A PDP model. *Cognitive Psychology, 44* (3), 297–337.

Goldrick, M., & Rapp, B. (2002). A restricted interaction account (RIA) of spoken word production: The best of both worlds. *Aphasiology, 16*(1–2), 20–55.

Gollan, T. H., & Ferreira, V. S. (2009). Should I stay or should I switch? A cost–benefit analysis of voluntary language switching in young and aging bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 640.

Gollan, T. H., & Silverberg, N. B. (2001). Tip-of-the-tongue states in Hebrew–English bilinguals. *Bilingualism: Language and Cognition, 4*(1), 63–83.

Gollan, T. H., Montoya, R. I., & Werner, G. A. (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology, 16*(4), 562.

Gollan, T. H., Montoya, R. I., Fennema-Notestine, C., & Morris, S. K. (2005). Bilingualism affects picture naming but not picture classification. *Memory & Cognition, 33*(7), 1220–1234.

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language, 58*(3), 787–814.

Gollan, T. H., Slattery, T. J., Goldenberg, D., van Assche, E., Duyck, W., & Rayner, K. (2011). Frequency drives lexical access in reading but not in speaking: The frequency-lag hypothesis. *Journal of Experimental Psychology: General, 140*, 186–209.

Gordon, J. K., & Dell, G. S. (2003). Learning to divide the labor: An account of deficits in light and heavy verb production. *Cognitive Science, 27*(1), 1–40.

Grainger, J., & Dijkstra, T. (1992). On the representation and use of language information in bilinguals. In *, Vol. 83. Advances in psychology* (pp. 207–220) (North-Holland).

Green, D. W. (1986). Control, activation, and resource: A framework and a model for the control of speech in bilinguals. *Brain and Language, 27*(2), 210–223.

Green, D. W. (1998a). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition, 1*(2), 67–81.
Green, D. W. (1998b). Schemas, tags and inhibition. *Bilingualism: Language and Cognition, 1*(2), 100–104.
Green, D. W., & Abutalebi, J. (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology, 25*(5), 515–530.
Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language, 38*(3), 313–338.
Gullifer, J. W., Kroll, J. F., & Dussias, P. E. (2013). When language switching has no apparent cost: Lexical access in sentence context. *Frontiers in Psychology, 4*, 278.
Halle, M., & Marantz, A. (1993). *Distributed morphology and the pieces of inflection*.
Halle, M., & Marantz, A. (1994). Some key features of distributed morphology. *MIT Working Papers in Linguistics, 21*(275), 88.
Hanulová, J., Davidson, D. J., & Indefrey, P. (2011). Where does the delay in L2 picture naming come from? Psycholinguistic and neurocognitive evidence on second language word production. *Language & Cognitive Processes, 26*(7), 902–934.
Harley, T. A. (1993). Phonological activation of semantic competitors during lexical access in speech production. *Language & Cognitive Processes, 8*(3), 291–309.
Hartsuiker, R. J., & Kolk, H. H. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology, 42*(2), 113–157.
Hartsuiker, R. J., & Pickering, M. J. (2008). Language integration in bilingual sentence production. *Acta Psychologica, 128*(3), 479–489.
Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science, 15*(6), 409–414.
Hatzidaki, A., Branigan, H. P., & Pickering, M. J. (2011). Co-activation of syntax in bilingual language production. *Cognitive Psychology, 62*(2), 123–150.
Hermans, D., Bongaerts, T., De Bot, K., & Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language and Cognition, 1*(3), 213–229.
Herring, J. R., Deuchar, M., Parafita Couto, M. C., & Moro Quintanilla, M. M. (2010). "I saw the madre": Evaluating predictions about codeswitched determinernoun sequences using Spanish–English and Welsh–English data. *Journal of Bilingual Education and Bilingualism, 13*, 553–573.
Hohenstein, J., Eisenberg, A., & Naigles, L. (2006). Is he floating across or crossing afloat? Cross-influence of L1 and L2 in Spanish-English bilingual adults. *Bilingualism, 9*(3), 249.
Hoshino, N., & Kroll, J. F. (2008). Cognate effects in picture naming: Does cross-language activation survive a change of script? *Cognition, 106*(1), 501–511.
Humphreys, G. W., Riddoch, M. J., & Quinlan, P. T. (1988). Cascade processes in picture identification. *Cognitive Neuropsychology, 5*(1), 67–104.
Ivanova, I., & Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta Psychologica, 127*(2), 277–288.
Jackson, G. M., Swainson, R., Cunnington, R., & Jackson, S. R. (2001). ERP correlates of executive control during repeated language switching. *Bilingualism: Language and Cognition, 4*(2), 169–178.
Jake, J. L., Myers-Scotton, C., & Gross, S. (2005). A response to MacSwan (2005): Keeping the matrix language. *Bilingualism: Language and Cognition, 8*(3), 271–276.
Jared, D., & Kroll, J. F. (2001). Do bilinguals activate phonological representations in one or both of their languages when naming words? *Journal of Memory and Language, 44* (1), 2–31.
Kambanaros, M., & Van Steenbrugge, W. (2006). Noun and verb processing in Greek–English bilingual individuals with anomic aphasia and the effect of instrumentality and verb–noun name relation. *Brain and Language, 97*(2), 162–177.
Kang, C., Fu, Y., Wu, J., Ma, F., Lu, C., & Guo, T. (2017). Short-term language switching training tunes the neural correlates of cognitive control in bilingual language production. *Human Brain Mapping, 38*(12), 5859–5870.
Kibort, A., & Corbett, G. G. (2008). *Grammatical Features Inventory*. University of Surrey. https://doi.org/10.15126/SMG.18.
Kirk, N. W., Kempe, V., Scott-Brown, K. C., Philipp, A., & Declerck, M. (2018). Can monolinguals be like bilinguals? Evidence from dialect switching. *Cognition, 170*, 164–178.
Kleinman, D., & Gollan, T. H. (2016). Speaking two languages for the price of one: Bypassing language control mechanisms via accessibility-driven switches. *Psychological Science, 27*(5), 700–714.
Koch, I., Prinz, W., & Allport, A. (2005). Involuntary retrieval in alphabet-arithmetic tasks: Task-mixing and task-switching costs. *Psychological Research, 69*(4), 252–261.
Kootstra, G. J., Van Hell, J. G., & Dijkstra, T. (2010). Syntactic alignment and shared word order in code-switched sentence production: Evidence from bilingual monologue and dialogue. *Journal of Memory and Language, 63*(2), 210–231.
Kroll, J. F., & Gollan. (2014). Speech planning in two languages: What bilinguals tell us about language production. In M. Miozzo, V. Ferreira, & M. Goldrick (Eds.), *The Oxford handbookof language production* (pp. 165–181). Oxford, England: Oxford University Press.
Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33*(2), 149–174.
Kroll, J. F., Bobb, S. C., & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism, 9*(2), 119.
Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The revised hierarchical model: A critical review and assessment. *Bilingualism: Language and Cognition, 13*(3), 373–381.
Kuzmina, E., Goral, M., Norvik, M., & Weekes, B. S. (2019). What influences language impairment in bilingual aphasia? A meta-analytic review. *Frontiers in Psychology, 10*, 445.
La Heij, W. (2005). Selection processes in monolingual and bilingual lexical access. In *Handbook of bilingualism: Psycholinguistic approaches* (pp. 289–307).
La Heij, W., Dirkx, J., & Kramer, P. (1990). Categorical interference and associative priming in picture naming. *British Journal of Psychology, 81*(4), 511–525.
Leboe, J. P., Whittlesea, B. W., & Milliken, B. (2005). Selective and nonselective transfer: Positive and negative priming in a multiple-task environment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1001.
Lehtonen, M., & Laine, M. (2003). How word frequency affects morphological processing in monolinguals and bilinguals. *Bilingualism: Language and Cognition, 6*(3), 213–225.
Lemaire, P., & Lecacheur, M. (2010). Strategy switch costs in arithmetic problem solving. *Memory & Cognition, 38*(3), 322–332.
Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1–38.
Linck, J. A., Kroll, J. F., & Sunderman, G. (2009). Losing access to the native language while immersed in a second language: Evidence for the role of inhibition in second-language learning. *Psychological Science, 20*(12), 1507–1515.
Loebell, H., & Bock, K. (2003). Structural priming across languages. *Linguistics, 41*(5, ISSU 387), 791–824.
Macnamara, J. T. (Ed.). (1967). *Problems of bilingualism*. Society for the Psychological Study of Social Issues.
Macnamara, J., Krauthammer, M., & Bolgar, M. (1968). Language switching in bilinguals as a function of stimulus and response uncertainty. *Journal of Experimental Psychology, 78*(2p1), 208.
MacSwan, J. (2005a). Codeswitching and generative grammar: A critique of the MLF model and some remarks on "modified minimalism". *Bilingualism: Language and Cognition, 8*(1), 1–22.
MacSwan, J. (2005b). Précis of a minimalist approach to intrasentential code switching. *Italian Journal of Linguistics, 17*(1), 55.
MacWhinney, B. (2005). Extending the competition model. *International Journal of Bilingualism, 9*(1), 69–84.
Mägiste, E. (1979). The competing language systems of the multilingual: A developmental study of decoding and encoding processes. *Journal of Memory and Language, 18*(1), 79.
Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: A reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 503.
Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language, 91* (5–2).
Marian, V., & Kaushanskaya, M. (2007). Cross-linguistic transfer and borrowing in bilinguals. *Applied PsychoLinguistics, 28*(2), 369.
Martin, C. D., Dering, B., Thomas, E. M., & Thierry, G. (2009). Brain potentials reveal semantic priming in both the "active" and the "non-attended" language of early bilinguals. *NeuroImage, 47*(1), 326–333.
Meijer, P. J., & Fox Tree, J. E. (2003). Building syntactic structures in speaking: A bilingual exploration. *Experimental Psychology, 50*(3), 184.
Meuter, R. F., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language, 40*(1), 25–40.
Miceli, G., Benvegnù, B., Capasso, R., & Caramazza, A. (1997). The independence of phonological and orthographic lexical forms: Evidence from aphasia. *Cognitive Neuropsychology, 14*(1), 35–69.
Midgley, K. J., Holcomb, P. J., Walter, J. B., & Grainger, J. (2008). An electrophysiological investigation of cross-language effects of orthographic neighborhood. *Brain Research, 1246*, 123–135.
Miozzo, M., & Caramazza, A. (1997). On knowing the auxiliary of a verb that cannot be named: Evidence for the independence of grammatical and phonological aspects of lexical knowledge. *Journal of Cognitive Neuropsychology, 9*, 160–166.
Miozzo, M., & Caramazza, A. (1998). Varieties of pure alexia: The case of failure to access graphemic representations. *Cognitive Neuropsychology, 15*(1–2), 203–238.
Miozzo, M., & Caramazza, A. (2003). When more is less: A counterintuitive effect of distractor frequency in the picture-word interference paradigm. *Journal of Experimental Psychology: General, 132*(2), 228.
Misra, M., Guo, T., Bobb, S. C., & Kroll, J. F. (2012). When bilinguals choose a single word to speak: Electrophysiological evidence for inhibition of the native language. *Journal of Memory and Language, 67*(1), 224–237.
Morford, J. P., Wilkinson, E., Villwock, A., Piñar, P., & Kroll, J. F. (2011). When deaf signers read English: Do written words activate their sign translations? *Cognition, 118*(2), 286–292.
Morsella, E., & Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 555.
Muysken, P. (2000). *Bilingual speech: A typology of codemixing*. Cambridge: Cambridge University Press.
Navarrete, E., & Costa, A. (2005). Phonological activation of ignored pictures: Further evidence for a cascade model of lexical access. *Journal of Memory and Language, 53* (3), 359–377.
Nicoladis, E. (2006). Cross-linguistic transfer in adjective-noun strings by preschool bilingual children. *Bilingualism, 9*(1), 15.
Nicoladis, E., Palmer, A., & Marentette, P. (2007). The role of type and token frequency in using past tense morphemes correctly. *Developmental Science, 10*(2), 237–254.

Pavlenko, A. (2009). Conceptual representation in the bilingual lexicon and second language vocabulary learning. *The Bilingual Mental Lexicon: Interdisciplinary Approaches*, 125–160.

Pfaff, C. W. (1979). Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language*, 291–318.

Philipp, A. M., Gade, M., & Koch, I. (2007). Inhibitory processes in language switching: Evidence from switching language-defined response sets. *European Journal of Cognitive Psychology, 19*(3), 395–416.

Pickering, M. J., Branigan, H. P., Cleland, A. A., & Stewart, A. J. (2000). Activation of syntactic information during language production. *Journal of Psycholinguistic Research, 29*(2), 205–216.

Poplack, S. (1980). "Sometimes I'll start a sentence in Spanish y termino en español": Toward a typology of code-switching. *Linguistics, 18*, 581–618.

Poulisse, N. (1999). *Slips of the tongue: Speech errors in first and second language production* (Vol. 20). John Benjamins Publishing.

Poulisse, N., & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics, 15*(1), 36–57.

Ransdell, S. E., & Fischler, I. (1987). Memory in a monolingual mode: When are bilinguals at a disadvantage? *Journal of Memory and Language, 26*(4), 392–405.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107*(3), 460.

Rapp, B., Benzing, L., & Caramazza, A. (1997). The autonomy of lexical orthography. *Cognitive Neuropsychology, 14*(1), 71–104.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition, 42*(1–3), 107–142.

Runnqvist, E., Strijkers, K., & Costa, A. (2014). Bilingual word access. In M. Miozzo, V. Ferreira, & M. Goldrick (Eds.), *The Oxford handbookof language production*. Oxford, England: Oxford University Press.

Sandoval, T. C., Gollan, T. H., Ferreira, V. S., & Salmon, D. P. (2010). What causes the bilingual disadvantage in verbal fluency? The dual-task analogy. *Bilingualism, 13*(2), 231.

Schneider, D. W., & Anderson, J. R. (2011). A memory-based model of Hick's law. *Cognitive Psychology, 62*(3), 193–222.

Schoonbaert, S., Hartsuiker, R. J., & Pickering, M. J. (2007). The representation of lexical and syntactic information in bilinguals: Evidence from syntactic priming. *Journal of Memory and Language, 56*(2), 153–171.

Schwieter, J. W., & Sunderman, G. (2008). Language switching in bilingual speech production: In search of the language-specific selection mechanism. *The Mental Lexicon, 3*(2), 214–238.

Sebastián-Gallés, N., & Bosch, L. (2005). *Phonology and bilingualism*. Oxford University Press.

Sebastián-Gallés, N., & Kroll, J. (2003). Phonology in bilingual language processing: Acquisition, perception, and production. *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, 279–318.

Sebba, M. (1998). A congruence approach to the syntax of codeswitching.International. *Journal of Bilingualism, 2*(1), 1–21.

Thierry, G., & Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences, 104*(30), 12530–12535.

Thomas, E. R. (2007). Phonological and phonetic characteristics of African American vernacular English. *Lang & Ling Compass, 1*(5), 450–475.

Thomas, M. S. C., & Allport, A. (2000). Language switching costs in bilingual visual word recognition. *Journal of Memory and Language, 43*(1), 44–46.

Ullman, M. T. (2001). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition, 4*(2), 105–122.

Van Hell, J. G., & De Groot, A. M. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition, 1*(3), 193–211.

Van Hell, J. G., Ormel, E., Van der Loop, J., & Hermans, D. (2009). Cross-language interaction in unimodal and bimodal bilinguals. In *16th Conference of the European Society for Cognitive Psychology. Cracow, Poland, September* (pp. 2–5).

Verhoef, K., Roelofs, A., & Chwilla, D. J. (2009). Role of inhibition in language switching: Evidence from event-related brain potentials in overt picture naming. *Cognition, 110*(1), 84–99.

Vigliocco, G., Antonini, T., & Garrett, M. F. (1997). Grammatical gender is on the tip of Italian tongues. *Psychological Science, 8*(4), 314–317.

Weinreich, U. (1953). *Languages in contact: Findings and problems*. New York: Linguistic Circle of New York.

Woolard, K. A. (2004). Codeswitching. *A Companion to Linguistic Anthropology*, 73–94.

Yeung, N., & Monsell, S. (2003). Switching between tasks of unequal familiarity: The role of stimulus-attribute and response-set selection. *Journal of Experimental Psychology: Human Perception and Performance, 29*(2), 455.