A neural-assembly based view on word production: the bilingual test case.

Kristof Strijkers

(Accepted manuscript – uncorrected version)

A Neural Assembly Based View on Word Production: The Bilingual Test Case.

Kristof Strijkers

Laboratoire Parole et Langage (LPL), UMR 7309, CNRS - Aix-Marseille Université, 13100, Aix-en-Provence, France.

Address for correspondence:
Kristof Strijkers
Centre National de la Recherche Scientifique (CNRS)
Laboratoire Parole et Langage (LPL) – Université d'Aix-Marseille
5, Av. Pasteur
13100 Aix-en-Provence, France
Kristof.Strijkers@gmail.com

Abstract

I will propose a tentative framework of how words in two languages could be organized in the cerebral cortex based on neural assembly theory, according to which neurons that fire synchronously are bound into large-scale distributed functional units (assemblies) which represent a mental event as a whole ('gestalt'). For language this means a word is engendered by widely distributed cell assemblies in which the different linguistic constituents are grouped together in action-perception circuits and become activated in parallel. In this article I will discuss the advantages of assembly coding over the traditional hierarchical convergence scheme. Recent evidence from language comprehension and production supporting the notion of neural assemblies representing words are discussed and a spatiotemporal model for word production based on this concept is presented. Finally, I will show how this view may be generalized to bilingualism, and explain mechanistically some key phenomena in the literature.

Keywords: Language production, bilingualism, neural coding, binding problems, cell assemblies.

In the last 30 years, much language research has been dedicated to understand where, when and how the building blocks of language (words) are represented in the brain. These endeavors have generated detailed brain language models on the spatial and/or temporal dynamics of word processing in the cerebral cortex (e.g., Pulvermuller, 1999; Friederici, 2011; Indefrey & Levelt, 2004; Hickok & Poeppel, 2007; Grainger & Holcomb, 2009; Pulvermuller & Fadiga, 2010; Indefrey, 2011; Hickok, 2012; Price, 2012; Dell et al., 2013; Hagoort, 2013). However, so far, most of this work has focused on monolinguals and much less theory and data relevant for the translation of such brain language models to bilingual populations are available. That is not to say that no neurophysiological research in bilingualism has been conducted. On the contrary, the availability of neuroscientific techniques has also generated a boom of neurocognitive investigations in the field of bilingualism. Notwithstanding, most of that research has focused on questions specific to bilingual populations such as second language learning, bilingual language control, the neuronal overlap of bilingual representations, etc. On itself this makes sense, since the primary interest of this research domain is to understand what makes the bilingual brain "special" and how bilinguals can cope with that "special status". As a consequence, bilingual researchers are often less interested in the general, basic dynamics underlying words in the brain (e.g., whether the system is serial, modular, interactive, parallel distributed, etc.) and rather generate models addressing the specific issues related to bilingualism. Here I will argue that knowledge concerning the basic spatiotemporal architecture of the bilingual system might help us to understand and constrain the bilingual-specific questions in a more detailed manner, and allow us to predict why and how the bilingual brain is different (or not) compared to the monolingual brain.

In this article I will attempt to propose a basic spatiotemporal model focusing on classical topics addressed in the monolingual literature such as processing dynamics (e.g., serial versus parallel) and representational organization (e.g., local versus distributed), and see whether those basic processing properties might shed light on the bilingual-specific questions which dominate research in the field. More concretely, I will first discuss the currently dominant spatiotemporal model of word production (Indefrey

& Levelt, 2004; Indefrey, 2011), and how this model is adopted to research in bilingualism (section 1). Next, I will focus on one particular (conceptual) problem with the underlying sequential hierarchical assumptions of the model when translated to neural coding schema of the brain. Specifically, I will discuss the computational problems of hierarchical and rate coding schemas and highlight an alternative conception of how to integrate neural processing with cognitive processing, namely assembly coding (section 2). Hereafter, I will extrapolate these coding issues to language production and propose a framework based on assembly theory with the objective of mechanistically linking word production components with neurophysiological phenomena (section 3). Finally, I will adopt this neural assembly model of language production to bilingualism and use different empirical patterns in bilingual speech production as a test case to explore whether and how a neural assembly based view can account for those findings (section 4). The reason to use bilingualism as test case is twofold: first, in general terms and as mentioned in the first paragraph, it exemplifies the importance for bilingual research to also take into account traditional questions in language such as processing dynamic and representational organization; second, using bilingualism as 'experimental variable' offers a particularly interesting contrast to understand and explore the notion of words as cell assemblies. This is because bilingualism concerns an extreme case where translation words have (near-) identical meaning which needs to be mapped onto (partially or fully) different phonology.

1. From monolingual spatiotemporal hierarchies to bilingual ones: The correlational approach.

When it comes to the cognitive architecture underpinning bilingual speech production, the most common practice is to adopt the traditional hierarchical structure of the psycholinguistic models of language production (e.g., Dell, 1986; Caramazza, 1997; Levelt et al., 1999). The assumption then is that bilinguals have a single conceptual system from where activity flows onto separate, but integrated lexicons (i.e., there are two lexicons, but they rely on the same system), followed by the retrieval of the phonemes in the first and/or second language, and ending with articulatory planning of an L1 or L2

utterance (e.g., Poulisse & Bongaerts, 1994; Green, 1998; Costa, 2005; Kroll & Tokowicz, 2005; Costa & Sebastian-Galles, 2014).

If we wish to translate this system to the level of the brain, and in particular to spatiotemporal dynamics, the most straightforward way is to extrapolate the spatial and temporal correlates identified for each hierarchical layer in monolingual models to the bilingual hierarchical structure. The most detailed and influential spatiotemporal model of word production components is that of Indefrey and Levelt (2004; henceforth I&L). In this model the authors propose (roughly) the following sequential activation dynamic from posterior towards anterior regions in the brain when preparing a speech utterance (see Fig. 1A): Lexico-semantic properties of an intended word for production are retrieved in the mid temporal gyrus (MTG) around 200 ms after stimulus (object) onset. Next, the word forms (lexical phonology or lexemes) linked to those lexico-semantic representations become activated after around 300 ms in the posterior MTG and the superior temporal gyrus (STG). Subsequently the brain enters in the encoding of the motor phonology of the word (syllabification and phonetics) around 400 ms from processing onset in the left inferior frontal gyrus (LIFG). Finally, from around 600 ms after processing onset until actual word production, the articulatory plans of the upcoming speech utterance are activated in the premotor and motor cortex.

Taking into account some key neurophysiological differences between first and second language processing (e.g., Perani et al., 1998; Indefrey, 2006; Abutalebi & Green, 2007; Runnqvist et al., 2011), a bilingual extension of this spatiotemporal map could look as follows (see Fig. 1B): Just as in the monolingual case, speech preparation starts with the retrieval of lexico-semantic knowledge in MTG and gradually moves ‘upwards’ through the system, over lexical phonology in the STG and motor phonology in the LIFG, until an articulatory command in the desired language is triggered in the motor cortex. Words in a bilingual’s first and second language would recruit largely overlapping brain systems which furthermore converge with those in place for monolingual speech production (e.g., Green, 2003; Runnqvist et al., 2012). Main spatial differences in cortical organization of L1 and L2 representations

(and with monolinguals) would be restricted to a more extended recruitment of the same neural tissue during L2 then L1 speech, in particular for left inferior frontal, frontocentral and prefrontal cortex associated with either motor phonology (e.g., Indefrey, 2006; Hanulova et al., 2011) or language control demands (e.g., Abutalebi & Green, 2007; Abutalebi et al., 2011; Branzi et al., 2015). In the temporal domain it is to be expected that the activation time-course of the brain regions is slower in L2 compared to L1 (e.g., Ivanova & Costa, 2008), at the very least from motor phonology onwards (e.g., Indefrey, 2006; Hanulova et al., 2011), but most likely already during lexico-semantic processing (e.g., Strijkers et al., 2010; 2013; Runnqvist et al., 2011).

This bilingual extension of the spatiotemporal correlates underpinning word production components, as displayed in Fig. 1B, is often used by researchers as an 'implicit guide' . The basic properties of this map are seldom subject of investigation, but rather used as an a-priori framework to help interpret spatial or temporal findings specific to a bilingual research question. Indeed, if the manner in which the brain organizes and processes language is sustained by a sequential hierarchical structure, then it seems justified that neuroscience research in bilingualism takes that basic architecture as an a-priori starting point and focuses on those issues specific to the bilingual case. This is because such structure relies on a fixed forward progression through processing hierarchies and is thus entirely defined by the anatomical connectivity between representational layers (e.g., van Rossum et al., 2002). Hence, although the speed and strength with which brain activity can cycle through the different hierarchies may depend on the type of representations (e.g., monolingual vs. bilingual word representations), the system itself will follow the same fixed anatomical structure and linguistic functionality regardless of context (e.g., monolingual vs. bilingual context). In sum, to the extent that the assumptions underpinning I&L's spatiotemporal map are correct, so will the bilingual adaptation of that spatiotemporal map be. However, currently, I&L's spatiotemporal map (2004) has been questioned on several grounds. For one, the temporally sequential and spatially local assumptions of the model lack sufficient empirical evidence. This is because (a) most temporal estimates stem from tasks which did not require natural and overt

speech behavior; and (b) the neural correlates elicited by speech production in many studies were fitted a-priori onto a serial architecture (and therefore do not allow for testing the model; e.g., Strijkers & Costa, 2011; 2016a). Relatedly, several current neurophysiological studies involving overt and immediate speech production, show time courses and cortical sources that do not always follow the above mentioned spatiotemporal division (for recent reviews see: e.g., Strijkers & Costa, 2011; 2016a; Munding et al., 2015). Given that these empirical issues with I&L's spatiotemporal map (2004) have been described recently in great detail, I will not discuss them further here. Instead I would like to focus on another, more conceptual problem with I&L's model based on the advances made in systems neuroscience.

As mentioned above, the underlying dynamics and structure defining I&L's model are serial processing coupled to a local hierarchical organization. Each of the linguistic components involved in word production have a specific activation time-course and are sustained by specialized cortical areas. In this manner, the core mental operations of speaking are ascribed to brain regions in an almost perfect one-to-one fashion. Despite the simplicity and transparency such direct correspondence between linguistic function and the brain offers, it is becoming increasingly clear that in many cases (if not all) mapping is one-to-many (a function is sustained by several, different brain areas) and many-to-one (different functions are sustained by the same brain region) (e.g., Smolensky & Legrande, 2006; Poeppel, 2012; Pulvermuller, 2013; Strijkers & Costa, 2016b). Consequently, in order to integrate psycholinguistic theory with neurophysiology, it is necessary to move beyond feedforward hierarchical seriality and start considering different mapping principles. Here I will attempt to do so for the spatiotemporal model of word production, and extend the resulting novel concepts to observations in bilingual speech production. To this end, I will start by making a brief excursion to systems neuroscience, in particular to the topic of neural coding, to get at the core of the debate concerning why a sequential hierarchical coding scheme fails to explain how the brain can bind and communicate complex representational knowledge. Subsequently, I will focus on an alternative view on neural coding which may be integrated with the psycholinguistics of word production in a new spatiotemporal model.

2. Which type of neural code can sustain the needs of language?

Adaptations of traditional language (in particular language production) models to spatiotemporal brain dynamics have mostly been done by assuming (implicitly or not) hierarchical anatomical convergence (i.e., one-to-one mapping), while considerations of other types of neural dynamics underpinning representational organization and communication in the cortex are rare. The goal of this section is to discuss certain problems with neural coding schemas which are (solely) based on the principle of hierarchical connectivity and demonstrate how these problems can be solved by taking into account time as an additional representational dimension (assembly coding). This is important because if basic neurophysiological principles of the brain cannot be captured by hierarchical coding alone, an integration of psycholinguistics with the brain that maintains such one-to-one mapping will likewise fail to capture the neural mechanics of language production[1].

One of the best-known (and adopted) principles by which the brain 'exchanges' cortical activity between 'input' and 'output' is based on the concept of hierarchical anatomical convergence, where integration of neural signals is the mechanism of communication (e.g., Barlow, 1972; Shadlen & Newsome, 1994; van Rossum et al., 2002). According to this concept, afferent signals ('input neurons') send their message by enhancing the average firing rate of the active neuronal population, and the receiving group of neurons ('response neurons') summate and integrate the post-synaptic potentials to trigger a corresponding response (action potential). This form of neural transmission is thus defined by the cortical connectivity between the axons of cells whose responses at a lower level of the hierarchy converge onto a common target cell(s) at a higher processing level. Neural coding through convergence is appealing because it follows the well-established understanding that the cortex is hierarchically organized

[1] Note, that this digression into systems neuroscience will necessarily be selective and simplified. Several of the concepts discussed can have different specific implementations and are often not mutually exclusive. The objective is merely to offer a more global, birds-eye view on how basic neurobiological brain properties can constrain psycholinguistic theory and processing.

into a collection of different areas (e.g., Hubel & Wiesel, 1962; Van Essen & Maunsell, 1983). This anatomical property coincides well with the classical notion in cognitive science that mental events and behavior are computed through a sequence of different and (more or less) independent processing steps, which each have a localizable substrate in the brain. The I&L (2004) model of word production, with its sequential processing hierarchies linked to local neural substrates, is an example of the assumption of neural communication through convergence in language.

Importantly, however, extensive theoretical and empirical research in systems neuroscience has identified several problems with hierarchical convergence coding. One problem is that it lacks flexibility. The anatomical connectedness from lower-level to higher-level neurons results in a deterministic and fixed system, while most of our cognitive abilities require flexibility in the routing of input to output in order to rapidly adapt to changing environments, intentions and tasks, and in order to learn novel information (e.g., Singer & Gray, 1995; Engel et al., 2001; Fries, 2005). Related is the problem of binding (e.g., Milner, 1974; von der Malsburg, 1985; Singer & Gray, 1995; Gray, 1999; Singer, 1999; 2013). In short, this problem refers to how our brain can code and combine distributed input to a coherent representation. Given that complex representations (e.g., objects) contain a multitude of features (e.g., color, shape, spatial location, movement, etc.), cortical input concerns distributed neural activity of different sensory features (e.g., Ballard et al., 1983; Felleman & Van Essen, 1991). How does our brain bind these features into a neural representation coding for the percept as a whole? In the traditional convergence scheme this is achieved by summating the responses of the distributed input onto a specific integration neuron(s) higher in the hierarchy to which the axons of all the input converge. While this neural structure may seem straightforward when considering the binding of features to objects in isolation, it becomes problematic when considering the many objects and nearly infinite feature combinations that exist in the world (i.e., the combinatorial problem; e.g., Milner, 1974; von der Malsburg, 1985; Singer & Gray, 1995). One would have to assume that each object as well as all its possible constellations are represented by a specific cell or small local group of cells (‘cardinal cells’;

e.g., Barlow, 1972). This would lead to an unrealistic number of neurons and an unacceptable number of connections to solve the binding problem.

An elegant modification of the hierarchical converging coding scheme in order to deal with the binding problem without leading to a combinatorial explosion of the required number of cells has been proposed in the form of population coding (e.g., Ballard et al., 1983; Felleman & Van Essen, 1991; Van Rossum et al., 2002). In this scheme the integration of distributed feature input is itself reflected by a distributed population of neurons along the cortical hierarchy (and to which that input is connected). Since in this case neurons can partake in more than one representation (what matters is a unique distributed pattern, not the individual 'nodes' making up the distributed pattern), convergence onto distributed populations largely reduces the number of cells required to represent different objects and their possible constellations. In this manner the combinatorial problem is solved while the notion of sequential processing across hierarchically organized cortical areas could be maintained (e.g., Ballard et al., 1983; Felleman & Van Essen, 1991). Although this constitutes an important shift from the 'single neuron doctrine' (Barlow, 1972) for complex representations to distributed neural populations, such system still lacks the flexibility that is essential to cognition (since the convergence remains solely governed by the structure of the anatomical connections along the hierarchy; e.g., Gray, 1999; Fries, 2005). Furthermore, an additional computational problem arises, the so-called 'superposition problem' (e.g., Milner, 1974; von der Malsburg, 1985; Singer & Gray, 1995). In essence this is an identification problem: If the system has to deal with two objects that share features (e.g., contour, color, luminance, motion, etc.), how does it know which features belong to which unique (distributed) representation (one for each object) higher in the hierarchy? Considering that everyday perception always involves multiple objects and backgrounds, and feature overlap amongst them is the common situation, superposition confronts the neural code with a real conceptual problem.

A theoretical solution to the superposition problem was based on a modification of Hebb's (1949) cell assembly concept (e.g., Milner, 1974; Braitenberg, 1978; Palm, 1980; Abeles, 1982; von der

Malsburg, 1985). Instead of binding feature input to a coherent representation by increasing the average firing rate between lower-level (input) and higher-level (output) neural populations, it was suggested that the synchronous firing of cortical cells in response to an object forms an assembly across the hierarchy that acts as an unitary entity. In other words, the relevant dimension for binding information in this scheme depends on temporal correlation where the specific timing of action potential discharges (rather than the average; e.g., Konig et al., 1996) between neural populations coding for the features and their combinations of a particular object synchronize on a milliseconds time-scale. Neurons coding for different objects would show no temporal correlation between them (out-of-sync) and thus different assemblies reflecting different complex representations are distinguished from one another by the temporal independence of their firing patterns. These processes are envisioned to occur in parallel, allowing for the formation of multiple cell assemblies simultaneously. Including temporal correlation of distributed activity between both proximate and distant cortical cells as an additional representational dimension is a crucial difference with population coding: a representation would not only be determined by the hierarchical anatomy of the brain, but also by its 'Gestalt' properties (crossing hierarchical anatomy). In this manner, assemblies are dynamical and unitary entities where neurons within the assembly interact more strongly with each other than those outside the assembly. As a consequence, multiple distributed neurons can be active for multiple objects at the same time, where the temporal dependencies between the activated neurons ensure the representational individuation between the objects; effectively resolving confusion and the superposition problem.

Strong empirical evidence for the notion of Hebbian-like neural assemblies and its underpinning synchronization were obtained in a series of multi-unit cell recording experiments demonstrating stimulus-specific synchronization of cortical activity in the visual cortex of cats (e.g., Eckhorn et al., 1988; Gray et al., 1989; Engel et al., 1991). For instance, it was shown that neural firing between columns of striate cortex was uncorrelated when bars in a display moved in opposite direction, but became correlated when the bars moved in the same direction and strongly synchronized when moving coherently

across the locations as a single object (e.g., Gray et al., 1989). In addition, it was demonstrated that such synchronized responses between brain cells resulted in enhanced synaptic sensitivity between them, while the absence of such temporal coherence weakened the synaptic strength (e.g., Tsodyks & Markram, 1997; Azouz & Gray, 2003). Taken together, the discussed data did not only show the existence of well-timed stimulus-dependent synchronous activity between cortical cells, but also that this synchrony followed Gestalt-like principles. These findings empirically fueled the idea that neurons in different areas of the cortex represent stimuli as functional units following Hebbian-like assembly principles, and were formulized in the 'binding-by-synchronization' framework of neural coding (Singer & Gray, 1995). Since then a wealth of evidence has confirmed and extended the notions of neural assemblies and binding by synchronization as key constructs to learn, represent and process information in the (human) brain (for focused reviews see: e.g., Buszaki & Draghun, 2004; Fries, 2009; Buszaki, 2010; Siegel et al., 2012; Singer, 2013). It is against this background I would like to re-evaluate I&L's spatiotemporal map of word production components. Specifically, in what follows I will discuss (a) why the neural coding problems discussed here also pose problems for sequential hierarchical models of word production, and (b) how a particular solution to these problems, namely assembly coding, might look when translated to the neurocognitive architecture of word production.

3. Neural assemblies in language and the spatiotemporal dynamics of word production.

3.1. The neural binding problem for words.

The above section served the purpose to highlight that at the neuronal level processing is not solely governed by anatomical connectivity from lower-level to higher-level brain areas, but involves precise temporal correlations between neural populations beyond the feedforward hierarchical structure of the brain. However, so far, I have mainly discussed the latter from the perspective of perception (where these issues have been most thoroughly discussed and investigated). Arrived at this point, a pressing

question is whether the neural coding problems are also of relevance for the topic of language. From the general perspective of brain-language integration it obviously does. That is, if neural synchronization is an important learning and processing principle of cortical function and organization, then the integration of cognition with that cortical function and organization cannot neglect a role for neural synchronization. Therefore, the notion of assembly coding has by now been embedded in neurobiological theories of cognitive abilities such as memory, attention, learning, consciousness and language (e.g., Mesulam, 1990; Fuster, 1995; Pulvermuller, 1999; 2005; 2013; Engel et al., 2001; Varela et al., 2001; Dehaene et al., 2006; Pulvermuller & Fadiga, 2010). Nevertheless, the manner in which assembly theory is implemented at a more specific level can take different forms. For instance, in language, several current neurocognitive models acknowledge Hebbian-like principles for learning linguistic knowledge (e.g., Hickok & Poeppel, 2007; Friederici, 2011; Hickok, 2012; Dell et al., 2013; Hagoort, 2013). That said, most of these models have focused on plotting linguistic function to advancements of the brain's neuroanatomy (e.g., Amunts & Catani, 2015) and are less explicit on the role of neural synchronization underpinning that functional anatomy (e.g., Friederici & Singer, 2015). In this manner, while witnessing a shift from segregated 'language processors' to interactive processing streams, a division of labor underpinning different linguistic properties of a word is maintained. From this perspective, the role attributed to neural synchronization (implicitly) is component specific, meaning that there are different assemblies for the different linguistic components of a word (e.g., lexico-semantics, syntax, phonology). In contrast to this, I will argue that implementing the notion of neural synchronization at the level of words requires the assumption that words, just as 'perceptual objects', are bound into unitary functional assemblies ('gestalts') (Pulvermuller, 1999; 2013). That is, distributed cell assemblies make up the cerebral fingerprint of a word where the semantic, lexical, phonological and articulatory properties bind together in action, perception and domain-general (integrating or 'switching/relaying') brain systems. I will try to defend this position by extrapolating the binding problem of neural coding described in the previous section to the binding of semantic and phonological knowledge in word learning.

In minimalistic terms, words concern the binding of sounds to meanings (and vice versa). Just as in the case of objects, the input will consist of a distributed pattern of different features (e.g., distributed semantic features in the case of production, and acoustic features in the case of perception). How does our brain bind these features together into a coherent and unique word representation? The easiest solution would be hierarchical convergence: simply integrate the activity of the different semantic features (e.g., 'round' and 'bouncing' for the concept *ball*) onto the specific cells to which those semantic features are connected 'upstream' in the hierarchy to trigger the corresponding phonological representation (e.g., /bal/). However, just as in the case of object perception, semantic or acoustic features can overlap between different words, giving rise to the combinatorial and superposition problems for the binding of linguistic input to words. An (extreme) example of this can be found in synonyms (e.g., *sofa – couch*, *baby – infant*, *job – work*, *pants – trousers*, etc.). In order to utter one of the synonyms, how does our brain link the strongly overlapping semantic features to the unique and intended word forms? A common assumption is that this can be achieved by attention or control mechanisms enhancing the saliency of the relevant combination (e.g., Olshausen et al., 1993). However, this does not solve the problem, it merely shifts it to the attentional mechanism: How does attention search for the different feature combinations and by which means is it 'attracted' to the relevant ones to enhance their saliency? Mappings (solely) based on hierarchical convergence between the input (semantics) and output (phonological form) would require unique and segregated semantic representations for each of the synonyms (even when having full overlap in meaning), otherwise the attentional system cannot enhance the target synonym's semantics to generate the correct phonological realization.

Importantly, this problem is not restricted to the special case of synonyms, but present for any word with overlap in semantic features (e.g., *dog – cat*). To be able to learn (and later produce) the link between the meaning of *dog* and its phonemes, without being in a continuous state of confusion with the overlapping sematic features of words such as *cat*, *animal* or *Dalmatian* (i.e., the superposition problem), those overlapping features need to converge to different and unique representations for every word

containing semantic overlap in order to map onto a specific word form[2]. Given the extensive feature overlap of various kinds that can exist between words (which furthermore increases exponentially when moving beyond the single word-level; e.g., combinatorial semantics, grammatical inflections, etc.), assuming for every potential overlap between words and all their possible constellations unique representations would lead to a combinatorial explosion of required representations and connections in the brain (i.e., the combinatorial problem). The same problems arise from the perception perspective (which for initial word learning may be of higher relevance). For example, after having learned a word such as *cat*, when perceiving for the first time the sounds in *cap*, how can a learner's brain link those sounds to the novel meaning of "clothing for the head" instead of (partially) mapping them to "animal" based on the already existing connections driven by the formal overlap of the other word? Put more generally, how can the system identify sound overlap between words (e.g., */bat/*, */bag/*, */bank/*, etc.) as bound to a unique solution, without assuming unique representations for the overlapping segments themselves (which would lead to a combinatorial explosion of representations and connections in the brain)?

As discussed in section 2, a prominent solution could be the temporal synchronization of the different features coding for a word. Temporal synchronization between lexico-semantics and phonology (and vice-versa) can ensure that overlapping semantic features (e.g., "mammal", "pet", "fur" and "four legs") of a target word (e.g., *cat*) are coupled to its particular lexical meaning without the need to assume distinct representations for the identical features in order to learn and avoid confusion with other words (e.g., *dog*). In this manner, by adding time as a representation differentiating principle, the same input-features (semantic or acoustic) can be 'recycled' for different words, resolving neural computation problems such as combinatorial explosion and superposition confusion. Consistent and coherent temporal

[2] Note that for comparable reasons (e.g., hyperonym problem) researchers such as Levelt assume the lexical network is nondecompositional in nature (e.g., Levelt, 1989; Levelt et al., 1999). While the notion of holistic lexical concepts mapping onto holistic lexical and phonological representations resolves confusion between semantic-to-phonology binding for a theory of lexical access, it does not resolve the problem itself; instead, the issue is shifted to the semantic-to-lexical concept mappings. More so, it concerns an extreme form of convergence (towards 'cardinal cells') in order to bind distributed semantic features to nondecompositional lexical concepts (strongly subject to both the combinatorial and superposition problems).

correlations between the semantic and phonological features of a word will result in the synaptic strengthening of their connections, while the absence of such temporal correlations will weaken connection strength. When sufficient learning has occurred, the correlated features form an assembly representing the word as a gestalt. Assembly-overlap between words would denote semantic relationship and generalization (e.g., *cat* and *dog* are *animals*), while assembly-dissociations would denote semantic saliency and individuation (e.g., a *cat meows* and a *dog barks*). And the same applies for phonology (e.g., *cat* and *cap* assemblies will overlap in the phonological features */ca/* but display distinct temporal correlations with their respective semantic connections due to the 'out-of-sync' differentiating phonological features */t/* and */p/*). In sum, if we acknowledge that neural synchronization is an important learning principle in the brain, then it seems reasonable to describe a word as a unitary functional assembly, since the co-occurrence of lexico-semantics and phonology of a specific word is maximal and unique to that word[3].

3.2. The spatiotemporal dynamics of words as neural assemblies.

What would the notion of words as unitary cell assemblies mean for the spatiotemporal dynamics and organization of a word in the brain? I will focus here on two distinctive properties argued to underpin word assemblies. A first property concerns the spatial organization. While in traditional one-to-one mapping models as that of I&L (2004) the different components making up a word are subserved each by local, specialized brain areas, an assembly view posits one-to-many distributed mappings between brain and linguistic function (e.g., Pulvermuller, 1999; 2013). Given that neural synchronization provides the

[3] Compare this for example with the co-occurrence in word combinations and sentences. For instance, although we mostly utter a noun in the presence of an article, the construction *'the ball'* will unlikely bind to a single assembly, since *'the'* can occur with any other noun as well. Thus co-occurrence is low in light of all the possible combinations (more synaptic weakening than strengthening between *'the'* and *'ball'*), contrary to the co-occurrence of the meaning features of ball and the phonology features of *ball* (continuous synaptic strengthening). Put differently, binding-by-synchronization and lasting assembly formation (at least in the case of storing memory traces as for words) does not mean that everything that is active in parallel at some point will bind to a functional unit. Rather, the binding must be coherent and consistent to increase synaptic strength. Given that the binding of semantics and phonology of a specific word meets those criteria, it seems parsimonious to perceive words as gestalts.

system with the capability to represent an intended target ('output') immediately through the binding with its input (e.g., Singer & Gray, 1995; Gray, 1999; Singer, 1999), neural assemblies underpinning words will concern large-scale representations where "low-level" input (e.g., auditory or visual features) and "high-level" output (e.g., semantics or articulation) form part of the word's inner assembly structure. As a consequence, peripheral systems such as sensory and motor cortex will play a role in the representational content of a word's assembly (note that this does not mean that Hebbian-like learning excludes the notion of disembodied, hub-like computations in more central areas of the network). Since these regions are well defined anatomically (e.g., somatotopic organization of the motor cortex), this allows us to identify word assemblies based on their differential local recruitment of sensory and motor cortex relevant to the linguistic properties of a word. A considerable amount of haemodynamic data in spoken and written comprehension has indeed confirmed topography-specific activations of sensorimotor systems linked to semantic and phonological knowledge of a word. For instance, color-related words activate the parahippocampal gyrus and fusiform gyrus, shape-related words trigger activity in the fusiform and dorsolateral prefrontal cortex, and action words activate the motor and premotor cortex in a somatotopic fashion (e.g., Hauk et al., 2004; Tettamanti et al., 2005; Simmons et al., 2007; Boulenger et al., 2009). Similar frontotemporal sensorimotor modulations have also been reported for phonological properties of a word. For instance, listening to speech sounds that are produced with the lips (e.g., */p/*) or the tongue (e.g., */t/*) activate the motor cortex somatotopically (e.g., Wilson et al., 2004; Pulvermuller et al., 2006), and disrupting those areas with transcranial magnetic stimulation results in behavioral dissociations when discriminating bilabial and alveolar speech sounds and words (e.g., D'Ausilio et al., 2009; Schomers et al., 2014).

A second property concerns the temporal dynamics of word activation. While in traditional models (e.g., I&L, 2004) activation is rather static and triggers the brain regions thought to subserve the different linguistic components sequentially or even in a serial fashion, cell assemblies are argued to display (at least) two functionally distinct activation time-courses: fast parallel *ignition* of the whole word

assembly followed by slower sequential *reverberations* in the whole assembly or in specific sub-parts of the assembly (e.g., the motor part of the assembly) (e.g., Pulvermuller, 1999; Pulvermuller et al., 2009; 2014; see also e.g., Fuster, 1995; Dehaene et al., 2006; Buszaki, 2010). Parallel *ignition* follows from the fact that the different linguistic components are embedded in the same word representation and when part of the assembly is activated (e.g., auditory or visually) it will quickly ignite and synchronize as a whole (taking into account that cortico-cortical fibers conduct action potentials with velocities around 10 m/s and high-frequency oscillations between different neurons occur on a ms-scale; e.g., Aboitiz et al., 1992; Konig et al., 1996). Several electrophysiological and MEG studies in language comprehension support this assembly property. A range of semantic, lexical, syntactical and phonological manipulations in single word processing paradigms elicited neurophysiological modulations between 100 and 200 ms after stimulus presentation (e.g., Menning et al., 2005; Hauk et al., 2006; Naatanen et al., 2007; Chanceaux et al., 2012; MacGreggor et al., 2012).

These findings are not easily accommodated by models portraying a hierarchical (and temporal) division of processing steps in word comprehension (e.g., Grainger & Holcomb, 2009; Friederici, 2011). Furthermore, the results call for a reevaluation of famous language-related ERP components such as the N400 and the P600 (e.g., Kutas & Hillyard, 1980; Osterhout & Holcomb, 1992; Hagoort et al., 1993). Given that the functionality associated with the N400 and P600, respectively lexico-semantics and syntax, could already be observed earlier in time, those brain responses are unlikely to reflect first-pass activation. Instead it has been suggested they reflect important second-level linguistic operations such as semantic integration and syntactic reprocessing (e.g., Pulvermuller et al., 2009). The latter proposal rhymes well with the *reverberatory* dynamic assigned to cell assembly 'behavior'. That is, after ignition assembly activation does not rapidly decay, but appears to be followed by slower sustained reverberations, which, importantly, seem to be affected by stimulus and task relevant properties (e.g., Fuster, 1995). In various domains, quick ignition has been linked to 'target' identification, and slower reverberations to second-order processes such as consolidation, reprocessing, decision-making and meta-

cognition (e.g., Mesulam, 1990; Fuster, 1995; Engel et al., 2001; Dehaene et al., 2006). In a similar vein, for language it has been proposed that those late, long-latency modulations (e.g., N400, P600) index this slow reverberatory activation within the ignited assembly and play a role in post-recognition operations such as semantic integration, grammatical inflections, syntactic framing and verbal working memory (e.g., Pulvermuller, 1999; 2002; Buszaki, 2010; Pulvermuller et al., 2014). Although evidence directly linking such fast versus slow(er) temporal dynamics to different functional processing states in language is still lacking (with the exception of computational simulations; e.g., Garagnani et al., 2008; Pulvermuller et al., 2014), recent discoveries in systems neuroscience showing that feedforward connections are more sensitive to higher-frequency oscillations while feedback connections are more sensitive to lower-frequency oscillations (e.g., Bosman et al., 2012; Bastos et al., 2012), seem to provide some support for this notion and a means to mechanistically implement it.

Interestingly, some recent neurophysiological observations in language production experiments display patterns which are likewise consistent with the above described spatial and temporal cell assembly properties. Several ERP studies on overt object naming have found that a range of psycholinguistic variables (e.g., semantic interference, lexical frequency, cognate-status and language membership) all elicited similar ERP modulations (P2-range) within 200 ms after picture onset (e.g., Costa et al., 2009; Strijkers et al., 2010; 2013). Especially the findings that cognate-status (whether or not translation words have phonological overlap) and language membership (L1 or L2 words) triggered such early ERP modulations were surprising since they are often associated with lexical phonology and motor phonology respectively (e.g., Indefrey, 2006; Christoffels et al., 2007). These data may thus suggest parallel activation of lexico-semantic and phonological properties in the course of speech planning, in line with the notion of word assembly ignition. Additional suggestive evidence for both the dynamical and organizational (at least in distributed terms) word assembly properties came from a recent meta-analysis of MEG studies on language production processes (Munding et al., 2015). Dividing source activations by brain region and linguistic component, the authors demonstrated that for several regions (anterior and

posterior ones) and linguistic functions (early and late in the traditional hierarchy) evidence for early activation (around 200 ms) has been reported. Along similar lines, Miozzo and colleagues (2014) performed a multiple-regression analysis on the neuromagnetic brain response elicited during overt object naming and found simultaneous activations around 150 ms after stimulus onset for lexico-semantic predictors in fronto-temporal regions and for a phonological predictor in the posterior MTG. Findings as these are difficult to reconcile with a serial and local spatiotemporal implementation of word production (e.g., I&L, 2004), but are predicted under an assembly scheme (although also interactive hierarchical models may still account for the findings; e.g., Strijkers et al., 2010).

More explicit evidence was reported recently in an anatomically constrained MEG study of overt object naming (Strijkers et al., under review). In that study the time-course of cortical area activations elicited by the lexical frequency of object names (a metric of lexico-semantic processes) was compared with the cortical area activations elicited by the particular articulatory movement (*lip vs. tongue*) required to utter different word-initial speech sounds (e.g., *m*onkey vs. *d*onkey) (a metric for acoustic and articulatory processes). Within 200 ms the lexico-semantic variable activated a fronto-temporal network (MTG and LIFG), consistent with previous heamodynamic sources and electrophysiological time-course reported for the lexical frequency effect (e.g., Graves et al., 2007; Strijkers et al., 2010). Crucially, in the same temporal window (160-240 ms) a simultaneous dependency of local and prediction-specific brain responses in the motor cortex and the STG with respect to the particular articulatory movement required to utter different initial speech sounds of a word was observed. The specific nature of the spatiotemporal pattern responding to a word's frequency in a fronto-temporal network and to a word's initial phoneme in a topography-dependent manner in the motor cortex (articulatory properties) and the STG (acoustic properties), provide compelling evidence that in the course of speech planning lexico-semantic and phonological-articulatory processes emerge rapidly together, drawing in parallel on temporal and frontal cortex.

Based on the issues described in neural coding, the data from language comprehension and the recent demonstrations in language production, Strijkers and Costa (2016a) extended the notion of neural assembly models to word production. The tentative framework that emerged from this exercise is schematically depicted in Figure 2. A word would be decoded at the level of the brain as a widely distributed inter-areal assembly involving the Perisylvian language areas as well as specific primary and/or secondary sensori and motor cortices depending on the particular meaning and phonology of words (e.g., parts of the temporo-occipital cortex for color or shape related words; for a detailed appreciation see: e.g., Pulvermuller, 2013). Binding through neural synchronization would couple posterior perception-related brain regions (yellow) with anterior action-related regions (red) reflecting integrated lexico-semantics of a word, and, similarly, sensori- (green) with motor (blue) regions reflecting integrated auditory-articulatory phonology of a word. As argued above, since words are based on consistent sound-meaning mappings in time, synchrony-dependent binding will eventually couple the sensori-motor networks of lexico-semantics and phonology into a coherent word assembly which rapidly ignites as a whole (within 200 ms). The more central regions (e.g., MTG, ATL, LIFG) are highly interconnected with the periphery of the system and may serve as convergence zones resulting in complex, hub-like representations containing 'abstracted' knowledge of word properties. Alternatively, they may function as relay stations ('switching neurons') instrumental in the routing of sensory to motor information (and vice versa) when no direct cortico-cortical connections between them exist (a hot and unresolved debate in the literature: e.g., Mahon & Caramazza, 2008; Pulvermuller, 2013). A third option (non-mutually exclusive with either of the above) is that some of these central, interconnected regions (especially in frontal cortex) serve domain-general functions relevant to language processing such as control, unification and sequencing (e.g., Federenko & Thompson-Schill, 2014; Hagoort, 2014). Note that a Hebbian-based perspective of word assemblies can be compatible with all of the above options.

In this manner, and similar to the Hebbian-based model in language comprehension (e.g., Pulvermuller, 1999; 2013), the assembly view on word production detailed here posits a widely

distributed spatial organization (across posterior and anterior brain regions) and two functionally distinct temporal processing dynamics (parallel ignition and sequential reverberation) underpinning our capacity to speak. This has several advantages over the traditional hierarchical sequential account that dominates the field. First, in general, it dispenses with neural computation problems of a one-to-one mapping between the psycholinguistic components of word production and the physiology of the brain (e.g., Singer & Gray, 1995; Pulvermuller, 1999; 2013; Smolensky & Legrande, 2006; Buszaki, 2010; Strijkers & Costa, 2016b). Second, more specifically, the assembly model may provide a better account for (and in fact a-priori predicts) the discussed current neurophysiological data in overt naming. The spatial property inherent to assemblies can easily accommodate the observation that word components (e.g., lexico-semantics or phonology) are not restricted to local brain regions (I&L, 2004), but recruit both posterior and anterior cortical areas, including sensori-motor systems (e.g., Miozzo et al., 2014; Munding et al., 2015; Strijkers et al., under review). The processing property of assemblies, integrating parallel distributed dynamics (ignition) and local sequential dynamics (reverberation), may explain the near-simultaneous activation of different word components (e.g., Costa et al., 2009; Strijkers et al., 2010; 2013; under review; Miozzo et al., 2014; Munding et al., 2015), as well as the sequential activation time courses reported in response to word production (e.g., Salmelin et al., 1994; Laganaro et al., 2009; Sahin et al., 2009; I&L, 2004; Munding et al., 2015); although the linguistic functionality attributed to those sequential responses will be different from the traditional models (linked to second-level, post-recognition operations rather than first-pass activation, see Figure 2). For example, possible functions of slow(er) reverberatory activity in word production may include semantic integration (across context and other words) and articulatory control (e.g., Strijkers & Costa, 2016a; Strijkers et al., under review). To explain this more concretely, take the following example: If a word is represented in the brain as a distributed cell assembly which ignites in parallel, then how are we capable of correctly uttering an intended word as */lap/* instead of its undesired mirror-image */pal/*? One possibility is that after the synchronous ignition, local reverberations within that particular assembly (e.g., motor cortex) ensure the timed articulation of the correct sequence /l/-/a/-/p/ instead of the incorrect sequence /p/-/a/-/l/ (which has a different internal

assembly structure linked to it due to non-overlapping lexico-semantics and thus a different reverberatory sequence or embedded 'synfire' chain; see Abeles, 1982; Pulvermuller, 2002).

In sum, although much empirical work remains to be done (in particular for language production), constraining models of brain-language integration with respect to its underlying neural code may offer more neurobiologically plausible and mechanistically explicit models compared to the currently available and more traditional neurocognitive theories of word production. The assembly based view presented in this section provides such explicit integration of word production with neural coding, resulting in a fundamentally different perspective on the cortical dynamics of speaking. If correct, this will have important consequences for research in bilingual speech production as well. A shift from perceiving words in the brain as a fixed system based on hierarchical sequential processing towards a flexible system based on temporal synchronization and non-hierarchical assemblies, will likewise involve a shift in theory and data interpretation of neuroscientific research in bilingualism. I will try to exemplify the latter in the final section of this paper, where I will adopt the assembly-based view to bilingual speech production and test how such model can explain, predict and extend neurophysiological observations of some well-known bilingual phenomena.

4. From monolingual assemblies to bilingual ones: The mechanistic approach.

A simplified schematic representation of the bilingual adaptation of the assembly model is presented in Figure 3 and follows the same color-coded functional-anatomical assembly structure as the monolingual version (Figure 2), but now one for each language[4]. In terms of the organizational principle, L1 and L2 assemblies have largely overlapping distributed lexico-semantic properties, suggesting that learning (Hebbian-based) words in the two languages rely on the synchronization and binding of the same

[4] The version of the model as depicted in Figure 3 concerns a standard situation of non-cognate words in early balanced bilinguals. Given that binding through synchronization is flexible, there is no single default network of words in the brain.

lexico-semantic features regardless of language. For example, in a Spanish-English bilingual, the English word *ball* and its Spanish translation *pelota* will synchronize the same visual features in occipito-temporal cortex (e.g., round, spherical, etc.) with the same action features in fronto-central cortex (e.g., rolling, bouncing, etc.), and consequently will rely on the same routing through the more central 'relay' or 'hub-like' brain regions (e.g., MTG, ATL, etc.). What differentiates the L1 and L2 words in this case is the distinct distribution of phonological and articulatory features in the STG (acoustics and lexical phonology), the LIFG and the frontocentral cortex (motor phonology and articulation). Returning to the *ball-pelota* example, it is evident that between both words there is a higher degree of distributional differences in the grouping of acoustic, phonological and articulatory features (/b/, /p/, /e/, /o/, /t/) than overlap (/a/, /l/). Note furthermore that binding through synchronizations offers an elegant explanation of how bilinguals can map the same meaning to different sounds (similarly as argued for solving the binding problem of objects and words discussed in sections 2 and 3). That is, the brain can differentiate translations based on the distinct temporal dependencies in synchronization of the same lexico-semantic features with the different phonological-articulatory features.

With regard to the dynamical principle, it is conceivable that the speed of both ignition and reverberation can differ between L1 and L2 assemblies, where a crucial determinant will be frequency of use. For example, more frequently activated assemblies (e.g., L1 words) will have more myelinated axons increasing conduction velocities and synchronization compared to less myelinated assemblies (e.g., L2 words) (e.g., Hartline & Colman, 2007). At a cognitive level, cell assemblies with higher myelination may translate to faster recognition (ignition) and, due to their higher saliency, display less reliance on reprocessing and control (reverberation). Of course, the extent of these velocity differences is sculpted by factors such as proficiency, AoA and frequency of use (e.g., Perani et al., 1998). Having broadly described the core model and its properties, I will now detail a few mechanistic accounts and novel predictions that can be derived from it with respect to cross-language activation, language membership and control, the cognate effect, and the spatiotemporal differences between speaking in a first or second language.

4.1. Cross-linguistic activations, language membership and control in bilingual cell assemblies.

An important topic in bilingual speech production is how bilinguals restrict their utterances to the desired language (i.e., language control; e.g., Green, 1998; Costa, 2005; Kroll et al., 2006; Abutalebi & Green, 2007). This question is driven by the well-established finding that words in both languages of a bilingual speaker become activated in the course of verbalization (e.g., Colome, 2001; Thierry & Wu, 2007). Given the latter, there should be a mechanism in place capable of selecting amongst the activated words those ones in the intended language. The nature and dynamics of such language control mechanism has been the subject of intense debate in the field. I will not discuss this extensive literature here, just point out to some interesting properties of the neural assembly view with respect to these issues.

A neural assembly model can (just as the more traditional hierarchical models) easily accommodate the dominant view of non-selective activation in bilingual speech production. Given that only partial stimulation of an assembly (e.g., certain sensory features) is sufficient to rapidly ignite the whole assembly (e.g., Pulvermuller, 1999; Pulvermuller et al., 2014), it naturally follows that words in the two languages of a bilingual become activated near-simultaneously (even when speaking in a monolingual context). This is because translation words are strongly overlapping assemblies: Stimulation of the lexico-semantic features (via an object for instance) are (nearly) identical in the two languages and will quickly synchronize with its learned and associated phonological and motor properties, which in the case of the bilingual are two differently distributed associations (see Figure 3). Thus, also under the assembly scheme, translation words become active in parallel and language control is required. Regardless the precise implementation (e.g., language-selective, non-selective, global inhibition, etc.), all mechanisms of control need to rely on some sort of classifiers indexing language membership. Typically, these classifiers are envisioned as lexicon-external language 'nodes' linked to a task schema (e.g., Green, 1998; see for comprehension: e.g., Dijkstra & Van Heuven, 2002). These language 'nodes' in turn are connected to their respective L1 or L2 words, allowing for regulation of lexical activation in function of

language membership. While an impressive amount of research has been conducted to understand the sources and nature of control (e.g., Abutalebi & Green, 2007; Green & Abutalebi, 2013), the notion of language membership has remained a rather abstract, computational curiosity. How are such nodes formed, what are the mechanics behind them and what does their neurobiological substrate look like? The neural assembly based view presented here may shed some light on these questions and offer a somewhat different perspective on the dynamics between language nodes communicating with lexical language tags to reflect the language membership of a word.

As discussed above, the most salient differences between translation words are phonological and articulatory in nature. From the assembly perspective this means that if we would take all first and second language words a bilingual knows and average them onto a brain template, there would be more distribution variability in the sensorimotor circuits reflecting the words' phonological and articulatory properties, than those related to the words' lexico-semantic properties (see Figure 4A). Given that this language-specific topography is quite consistently and repeatedly linked to one context (L1 production) or another (L2 production), supraordinate assemblies reflecting this language-specificity could emerge through a similar binding-by-synchronization principle described for the formation of word assemblies (although arguably on a longer time-scale). Upon this idea, language membership would be (1) an emergent phenomenon bound to a context-consistent variability in assembly distributions, (2) lexicon-internal as a direct consequence of word assembly structure, and (3) located especially in superior temporal and frontocentral cortex where variations in acoustic-articulatory properties between words in a first and second language will be maximal.

While the latter offers a mechanistic account and neurobiological substrate for 'language tags' in psycholinguistic models of bilingualism, it may still be required to have language nodes in small, localized groups of 'command neurons' from where top-down control can identify and modulate lexical representations with the appropriate (or inappropriate) 'language tags'. This remains indeed a viable option, but not a necessary one under an assembly scheme. Top-down modulatory activity could also be

instantiated from the language-specific supraordinate assemblies themselves (e.g., Engel et al., 2001). External and internal factors, such as auditory input in conversation, a face, an instruction or intention, can activate one of the language member assemblies in a proactive fashion, thereby enhancing the baseline activity of the word assemblies overlapping with the supraordinate assembly (e.g., Strijkers et al., 2011; 2015). The latter causes an advantage for one category of responses (a particular language in this case) over others, biasing competition in favor of the intended representations (e.g., Desimone & Duncan, 1995). In this sense, language control can be seen as a lexical-intrinsic source of contextual modulations. Of course, the full top-down network will include activity in systems involved in goal definition, action planning, working memory and selective attention (located in prefrontal, parietal and limbic areas; e.g., Corbetta & Shulman, 2002), corresponding to the language control network described in bilingualism (e.g., Abutalebi & Green, 2007). However, it would not be necessary to define abstracted notions of language membership in those systems themselves. Instead, through temporal binding, the formed language membership assemblies can synchronize with the executive control system forming a temporally large-scale network in function of the particular task, intention or context. Thus the same supraordinate language membership assemblies would function as 'language tags' (when bound with the words) and 'language nodes' (when bound with executive control systems).

4.2. Cognate assemblies.

Cognates constitute a special class of translation words with partial or full overlap in phonology (e.g., the Spanish-English pair *guitarra-guitar*). A typical observation (i.e., the cognate effect) is that cognates are named faster compared to non-cognates (e.g., the Spanish-English pair *tambor-drum*) (e.g., Costa et al., 2000). The bilingual field has extensively made use of these type of words to investigate issues such as cross-language activation and language control (e.g., Kroll et al., 2006), even though the exact nature and locus of the cognate effect remains debated. A straightforward account is that the processing advantage for cognates stems from the phonological overlap, resulting in less cross-linguistic

competition at the sublexical phonological level (e.g., Christoffels et al., 2007). However, the observations that cognate-status of a word can affect lexical factors and elicits early ERP effects suggests that cognates may already be differently represented compared to non-cognates at the earliest stages of lexical access (e.g., Costa et al., 2000; Strijkers et al., 2010). Yet another account posits that cognates may have more word-specific conceptual overlap than non-cognates, which could explain their differential sensitivity in semantic word association (e.g., Van Hell & De Groot, 1998).

The assembly-based view presented here provides a straightforward answer about the origin of cognates and the nature of the cognate effect. Following the binding-by-synchrony principle, the higher the overlap in phonology-articulatory features between translation words, the higher the overlap in L1 and L2 assemblies (Figure 4B). In this manner, non-cognates will have relatively small overlap in the sensorimotor circuits of a word's phonology (lexical and motor), non-identical cognates will have considerably more overlap and identical cognates will have almost full overlap (though still some smaller variations in acoustics and phonetics, depending on the degree of language similarity). In this context, the locus of the cognate effect is phonological in nature (e.g., Christoffels et al., 2007) and should be situated in the sensorimotor circuits binding the acoustic and motor properties of a word's phonology. The fact that haemodynamic studies of picture naming involving a cognate - non-cognate contrast report increased activity in frontocentral areas and STG (e.g., De Bleser et al., 2003; Palomar-Garcia et al., 2015), meshes well with the spatial predictions put forward by the assembly model. Furthermore, since cell assemblies operate as functional units where all properties pertaining to that unit are activated in parallel, the model also explains why the cognate effect occurs early (200 ms) and with a similar time-course as found for lexico-semantic manipulations (e.g., Strijkers et al., 2010). Finally, it can equally account for 'interactions' of cognate status with lexical factors and why semantic word associations of cognates are easier and more often translations of one another compared to non-cognates (e.g., Van Hell & De Groot, 1998). Put simply, this is because in the assembly-based model there is no hierarchical division between linguistic components to which the different effects should be fitted. In fact, there is no single localizable

lexicon that communicates with phonology higher in hierarchy and word semantics lower in the hierarchy. What is 'lexical' in the model is the word assembly, with its bound lexico-semantics and phonology, as a whole.

To conclude this subsection I want to point out that representational differences in L1 and L2 are not necessarily solely observable in the sensorimotor circuits underpinning phonological and articulatory word properties (as focused on so far). Although for the majority of situations, words and language-combinations, the most salient differences will be phonology-related, under specific conditions and for certain classes of words, distributional dissociations between L1 and L2 in the lexico-semantic 'part' of the assembly are possible (Figure 4C). For instance, abstract and emotion words may vary in their lexico-semantic features and associations between the first and second language of a bilingual (e.g., Hsu et al., 2015). In this case, certain distributional differences of the lexico-semantic circuit between L1 and L2 assemblies in brain areas such as the orbitofrontal cortex, the hippocampus and subcortical structures are expected. And even for concrete nouns, language-specific labeling of colors, objects, motion and cultural knowledge may induce processing differences between a bilingual's first and second language (e.g., Thierry et al., 2009; Ellis et al., 2015). From an assembly perspective these differences emerge because of variations between L1 and L2 assembly binding for language-combinations where such perceptual and cultural differences exist (e.g., distributional differences between L1 and L2 in the binding of color properties in occipito-temporal cortex and their corresponding phonology – one vs. two verbal labels for a color – in STG and motor cortex; color-example based on Thierry et al., 2009). In this manner, an assembly model would not only predict such differences, but may be able to explain the mechanics behind them at the neuronal level.

4.3. The language effect and the spatiotemporal differences between first and second language words.

Finally, I will provide a concrete example on how the cell assembly notion as implemented here could explain apparently inconsistent findings (from the traditional hierarchical perspective) concerning the origin of the language effect. The language effect refers to the observation that bilinguals are faster in their first compared to their second language (even when being highly proficient in both languages from birth; e.g., Ivanova & Costa, 2008). While that observation itself is not so surprising, addressing the question why and where in the brain these differences occur has been proven less straightforward. Based on haemodynamic evidence, there is general consensus by now that L1 and L2 naming recruit the same neural tissue, and the only clear differences are a more extended recruitment of prefrontal, inferior frontal (especially the LIFG) and frontocentral areas of the brain (e.g., Indefrey, 2006; Abutalebi, 2008; see also Sebastian et al., 2011 for evidence including certain temporo-parietal regions as well). These observations have been linked to more effortful processing of motor phonology (in particular syllabification) in L2 compared to L1 (for inferior frontal and frontocentral areas; e.g., Indefrey, 2006; Hanulova et al., 2011) and enhanced language control during L2 speech (for the prefrontal and inferior frontal differences; e.g., Abutalebi & Green, 2007; Abutalebi, 2008). However, when L1 and L2 object naming are compared with time-sensitive electrophysiological measures, the ERPs dissociate between languages very early on (within 200 ms) and cause similar modulations (P2) as found for lexico-semantic processes in word production (e.g., Strijkers et al., 2010; 2013). The combination of the haemodynamic and the electrophysiological findings are problematic for a hierarchical sequential model. This is because, when following the spatiotemporal estimates of word production components as detailed in I&L (2004), the two measures suggests different components to be responsible for the language effect. While the spatial data indicate a late origin of the language effect linked to motor phonology (e.g., LIFG), the temporal data indicate an early, lexico-semantic origin of the language effect (in principle associated with temporal brain areas in I&L).

Interestingly, the assembly model presented here suggests that the spatial and temporal differences between L1 and L2 production are not necessarily contradictive. In an assembly-based view

on words in the bilingual brain there are no functional differences in the activation time course of posterior versus anterior brain regions. This is because neural populations linked to the 'input' (e.g., lexico-semantics) and those associated with the 'output' (e.g., phonology and articulation) synchronize on an ms-scale. Hence, an assembly model is not bound to the assumption that the frontal area activation occurs late in the speech preparation process, but instead predicts the fast and near-simultaneous ignition of frontotemporal circuits underpinning words. In this manner, the observation of early ERP modulations between L1 and L2 translations in combination with especially frontal brain modulations are predicted in an assembly framework of bilingual word production. As mentioned in the two previous subsections, on average and for concrete nouns, the most salient differences between L1 and L2 word assemblies will be housed in the brain regions containing the acoustic, phonological and articulatory features of the assembly (i.e., STG, motor and premotor cortex, and LIFG). These distributional differences between languages in brain regions responsible for phonology and articulation together with the dynamical principle of assembly ignition (>200ms) can therefore easily explain and reconcile the temporal findings with EEG and spatial findings with fMRI (and PET) for the language effect.

One apparent caveat with the above assembly explanation for the spatiotemporal differences between translation words in bilingual speech production needs to be addressed. One may object that an assembly view can only partially account for spatiotemporal observations between first and second language processing, since the model should also predict spatial differences between L1 and L2 words in superior temporal brain regions associated with acoustic and phonological properties. The latter does not seem supported by the available haemodynamic data which highlights that differences between L1 and L2 production are especially located in frontal areas of the brain (e.g., Indefrey, 2006; Abutalebi & Green, 2007; Abutalebi, 2008; but see Sebastian et al., 2011). However, such conclusion may be premature at present for two reasons: first, stronger activation for L2 naming compared to L1 naming in the frontal cortex is not only related to motor phonology, but has also been associated with enhanced cognitive control during L2 speech production (e.g., Abutalebi & Green, 2007). As a consequence, it may be easier

to pick up language differences in the frontal cortex because of its sensitivity to two potential forces modulating L1-L2 activation (motor phonology and language control). This seems particularly relevant for the LIFG, which could be involved in both types of operations. Second, most of the haemodynamic studies did not specifically control for or manipulate distributional differences in the amount of acoustic-articulatory feature overlap between L1 and L2 words. In the bilingual word assembly scheme this is an essential factor by which language-specific topographies in auditory and motor cortex (and perhaps LIFG) can occur. This argumentation is supported by data of electrocortical stimulation mapping (ESM) studies in bilingual patients. ESM is a cortical localization technique mapping important motor and language areas through direct electrical stimulation of the cerebral cortex in pre-surgical patients in order to avoid unnecessary functional damage. This is relevant in the present context because ESM is highly sensitive to distributional differences between conditions (e.g., languages). Interestingly, a number of ESM studies in bilingual patients comparing L1 versus L2 naming report local distributional language differences in both the frontocentral cortex and the STG (e.g., Roux & Tremoulet, 2002; Lucas et al., 2004; Cervenka et al., 2011). These data, together with the previously discussed haemodynamic and electrophysiological results, are remarkably consistent with the full scope of predictions of the bilingual assembly model, which at present seems to offer the most parsimonious account concerning the spatiotemporal differences between first and second language production and the behavioral language effect arising from these.

5. Conclusion

In this article I have argued that general principles such as cortical organization and activation dynamics of words in the brain are of importance to explain bilingual-specific topics, if we question the typical hierarchical sequential structure underpinning brain-language integration that dominates the field of language production. I have argued that such one-to-one mapping structure will suffer from computational problems such as combinatorial explosion and superposition confusion when binding the

lexico-semantic and phonological knowledge of a word. To dispense with these neural binding problems, I have suggested adopting the principles of temporal correlation and assembly formation to word production. Spatial organization of words in the brain would rely on one-to-many mappings, resulting in inter-areal distributed functional circuits that represent words as gestalts. The temporal dynamics underpinning word production would rely on both parallel distributed and local sequential processing, where ignition reflects word activation and reverberation reflects secondary language operations such as articulatory control. Importantly, this structural shift from a fixed hierarchical to a dynamical non-hierarchical system for word production equally means a shift in how to mechanistically explain and advance in bilingual-specific research questions. Extending the assembly based view on word production to two languages (1) offered a novel and neurobiologically plausible account of 'language nodes' and how these cortical substrates can ensure language control, (2) predicted the differences in cortical organization between cognate and non-cognate words and how that organization can affect processing early on, and finally (3) explained seemingly inconsistent findings (from a hierarchical perspective) between the time-course and brain sources underpinning word production differences between a bilingual's first and second language.

References

Abeles, M. (1982). *Local cortical circuits* (p. 102). Springer, Berlin.

Aboitiz, F., Scheibel, A. B., Fisher, R. S., & Zaidel, E. (1992). Fiber composition of the human corpus callosum. *Brain research*, *598*(1), 143-153. doi:10.1016/0006-8993(92)90178-C

Abutalebi, J. (2008). Neural aspects of second language representation and language control. *Acta psychologica*, *128*(3), 466-478. doi:10.1016/j.actpsy.2008.03.014

Abutalebi, J., & Green, D. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of neurolinguistics*, *20*(3), 242-275. doi:10.1016/j.jneuroling.2006.10.003

Abutalebi, J., et al. (2011). Bilingualism tunes the anterior cingulate cortex for conflict monitoring. *Cerebral Cortex*, bhr287. doi: 10.1093/cercor/bhr287

Amunts, K., & Catani, M. (2015). Cytoarchitectonics, receptor architectonics, and network topology of language. *The Cognitive Neurosciences, edn 5*. Edited by Gazzaniga MS. Cambridge, MA: MIT Press.

Azouz, R., & Gray, C. M. (2003). Adaptive coincidence detection and dynamic gain control in visual cortical neurons in vivo. *Neuron*, *37*(3), 513-523. doi:10.1016/S0896-6273(02)01186-8

Ballard, D. H., Hinton, G. E., & Sejnowski, T. J. (1983). Parallel visual computation. *Nature*, *306*(5938), 21-26. doi:10.1038/306021a0

Barlow, H.B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* **1** 371–94.

Bastos, A. M., et al. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695-711. doi:10.1016/j.neuron.2012.10.038

Bosman, C. A., et al. (2012). Attentional stimulus selection through selective synchronization between monkey visual areas. *Neuron*, *75*(5), 875-888. doi:10.1016/j.neuron.2012.06.037

Boulenger, V., Hauk, O., & Pulvermüller, F. (2009). Grasping ideas with the motor system: semantic somatotopy in idiom comprehension. *Cerebral Cortex*, *19*, 1905-1914. doi: 10.1093/cercor/bhn217

Braitenberg, V. (1978). Cell assemblies in the cerebral cortex. In: *Theoretical approaches to complex systems. (Lecture notes in mathematics, vol. 21),* ed. R. Heim & G. Palm. Springer.

Branzi, F. M., Della Rosa, P. A., Canini, M., Costa, A., & Abutalebi, J. (2015). Language Control in Bilinguals: Monitoring and Response Selection. *Cerebral Cortex*, bhv052. doi: 10.1093/cercor/bhv052

Bressler, S. L., Coppola, R., & Nakamura, R. (1993). Episodic multiregional cortical coherence at multiple frequencies during visual task performance. *Nature*, *366*(6451), 153-156. http://dx.doi.org/10.1038/366153a0

Buzsáki, G. (2010). Neural syntax: cell assemblies, synapsembles, and readers. *Neuron*, *68*(3), 362-385. doi:10.1016/j.neuron.2010.09.023

Buzsaki, G., & Draghun, A. (2004). Neuronal oscillators in cortical networks. *Science, 304*, 1926-1929. DOI: 10.1126/science.1099745

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology,* 14, 177–208. DOI:10.1080/026432997381664

Cervenka, M. C., Boatman-Reich, D., Ward, J., Franaszczuk, P. J., & Crone, N. (2011). Language mapping in multilingual patients: electrocorticography and cortical stimulation during naming. *Frontiers in human neuroscience*, *5*, 13. http://dx.doi.org/10.3389/fnhum.2011.00013

Chanceaux, M., Vitu, F., Bendahman, L., Thorpe, S & Grainger, J. (2012). Word processing speed in peripheral vision measured with a saccadic choice task. *Vision Research,* 56, 10–19. doi:10.1016/j.visres.2012.01.014

Christoffels, I.K., Firk, C., & Schiller, N.O. (2007). Bilingual language control: an event-related brain potential study. *Brain Research,* 1147, 192–208. doi:10.1016/j.brainres.2007.01.137

Colomé, A. (2001). Lexical activation in bilinguals' speech production: language-specific or language-independent? *Journal of Memory and Language*, 45, 721-736. doi:10.1006/jmla.2001.2793

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, *3*(3), 201-215. doi:10.1038/nrn755

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., 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. http://dx.doi.org/10.1037/0278-7393.26.5.1283

Costa, A., & Sebastián-Gallés, N. (2014). How does the bilingual experience sculpt the brain?. *Nature Reviews Neuroscience*, *15*(5), 336-345. doi:10.1038/nrn3709

Costa, A., Strijkers, K., Martin, C., & Thierry, G. (2009). The time-course of word retrieval revealed by event-related brain potentials during overt speech. *Proceedings of the National Academy of Sciences*, 106, 21442-21446. doi:10.1073/pnas.0908921106

D'Ausilio, A., et al. (2009). The motor somatotopy of speech perception. *Current Biology,* 19, 381–385. doi:10.1016/j.cub.2009.01.017

De Bleser, R., et al. (2003). The organisation of the bilingual lexicon: A PET study. *Journal of Neurolinguistics*, *16*(4), 439-456. doi:10.1016/S0911-6044(03)00022-8

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, 10(5), 204-211. doi:10.1016/j.tics.2006.03.007

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., & Coslett, H. B. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition*, 128, 380-396. doi:10.1016/j.cognition.2013.05.007

Dell, G.S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review,* 93, 283–321. http://dx.doi.org/10.1037/0033-295X.93.3.283

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition, 42,* 287–314. doi:10.1016/0010-0277(92)90046-K

Dijkstra, T., & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(03), 175-197. http://dx.doi.org/10.1017/S1366728902003012

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*(1), 193-222. DOI: 10.1146/annurev.ne.18.030195.001205

Eckhorn, R., et al. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex? *Biological cybernetics*, *60*(2), 121-130. 10.1007/BF00202899

Ellis, C., et al. (2015). Language and culture modulate online semantic processing. *Social cognitive and affective neuroscience*, nsv028. doi: 10.1093/scan/nsv028

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience, 2,* 704-716. doi:10.1038/35094565

Engel, A. K., König, P., Kreiter, A. K., & Singer, W. (1991). Interhemispheric synchronization of oscillatory neuronal responses in cat visual cortex. *Science*, *252*(5009), 1177-1179. DOI: 10.1126/science.252.5009.1177

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in cognitive sciences*, *18*, 120-126. doi:10.1016/j.tics.2013.12.006

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, *1*(1), 1-47. doi: 10.1093/cercor/1.1.1

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, *91*(4), 1357-1392. DOI: 10.1152/physrev.00006.2011

Friederici, A. D., & Singer, W. (2015). Grounding language processing on basic neurophysiological principles. *Trends in cognitive sciences*. doi:10.1016/j.tics.2015.03.012

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in cognitive sciences*, *9*(10), 474-480. doi:10.1016/j.tics.2005.08.011

Fries, P. (2009). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual review of neuroscience*, *32*, 209-224. DOI: 10.1146/annurev.neuro.051508.135603

Fuster, J. M. (1995). Memory in the cerebral cortex. MIT press, Cambridge, MA.

Goldrick, M., Dell, G. S., Kroll, J., & Rapp, B. (2009). Sequential information processing and limited interaction in language production. *Science (online letters).*

Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: on the time-course of component processes in visual word recognition. *Language and Linguistic Compass*, 3, 128–156. DOI: 10.1111/j.1749-818X.2008.00121.x

Graves, W.W., Grabowski, T.J., Mehta, S., & Gordon, J.K. (2007). A neural signature of phonological access: distinguishing the effects of word frequency from familiarity and length in overt picture naming. *Journal of Cognitive Neuroscience,* 19, 617–631. doi:10.1162/jocn.2007.19.4.617

Gray, C. M. (1999). The temporal correlation hypothesis of visual feature integration: still alive and well. *Neuron*, *24*(1), 31-47. doi:10.1016/S0896-6273(00)80820-X

Gray, C. M., König, P., Engel, A. K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, *338*(6213), 334-337. doi:10.1038/338334a0

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, 1, 67-81. http://dx.doi.org/10.1017/S1366728998000133

Green, D. W. (2003). The neural basis of the lexicon and the grammar in L2 acquisition. *The interface between syntax and the lexicon in second language acquisition*, 197-208.

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

Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Frontiers in psychology*, *4*. doi:10.3389/fpsyg.2013.00416.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and cognitive processes*, *8*(4), 439-483. DOI:10.1080/01690969308407585

Hanulová, J., Davidson, D.J., & Indefrey, P. (2011). Where does the delay in L2 picture naming come from? Neurocognitive evidence on second language word production. *Language and Cognitive Processes,* 26, 902–934. DOI:10.1080/01690965.2010.509946

Hartline, D. K., & Colman, D. R. (2007). Rapid conduction and the evolution of giant axons and myelinated fibers. *Current Biology*, *17*(1), R29-R35. doi:10.1016/j.cub.2006.11.042

Hauk, O., et al. (2006). [Q:] When would you prefer a SOSSAGE to a SAUSAGE? [A:] At about 100 msec ERP correlates of orthographic typicality and lexicality in written word recognition. *Journal of Cognitive Neuroscience*, 18, 818–832. 10.1162/jocn.2006.18.5.818

Hauk, O. et al. (2004). Somatotopic representation of action words in the motor and premotor cortex. *Neuron* 41, 301–307. doi:10.1016/S0896-6273(03)00838-9

Hebb, D. O. (1949). *The organization of behavior. A neurophysiological theory.* Wiley.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*, 135-145. doi:10.1038/nrn3158

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience,* 8, 393–402. doi:10.1038/nrn2113

Hsu, C. T., Jacobs, A. M., & Conrad, M. (2015). Can Harry Potter still put a spell on us in a second language? An fMRI study on reading emotion-laden literature in late bilinguals. *Cortex*, *63*, 282-295. doi:10.1016/j.cortex.2014.09.002

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, 2, 255. doi: 10.3389/fpsyg.2011.00255

Indefrey, P., & Levelt, W.J.M. (2004). The spatial and temporal signatures of word production components. *Cognition,* 92, 101–144. doi:10.1016/j.cognition.2002.06.001

Indefrey, P. (2006). A Meta-analysis of Hemodynamic Studies on First and Second Language Processing: Which Suggested Differences Can We Trust and What Do They Mean? *Language Learning*, *56*(s1), 279-304. DOI: 10.1111/j.1467-9922.2006.00365.x

Ivanova, I., & Costa, A. (2008). Does bilingualism hamper lexical access in speech production?. *Acta psychologica*, *127*(2), 277-288. doi:10.1016/j.actpsy.2007.06.003

König, P., Engel, A. K., & Singer, W. (1996). Integrator or coincidence detector? The role of the cortical neuron revisited. *Trends in neurosciences*, *19*(4), 130-137. doi:10.1016/S0166-2236(96)80019-1

Kroll, J. F., & Tokowicz, N. (2005). Models of bilingual representation and processing. *Handbook of bilingualism: Psycholinguistic approaches*, 531-553.

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: Language and Cognition*, 9, 119–135. http://dx.doi.org/10.1017/S1366728906002483

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203-205. http://dx.doi.org/10.1126/science.7350657

Levelt, W.J.M., Roelofs, A., & Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral Brain Sciences,* 22, 1–38.

Lucas, T. H., McKhann, G. M., & Ojemann, G. A. (2004). Functional separation of languages in the bilingual brain: a comparison of electrical stimulation language mapping in 25 bilingual patients and 117 monolingual control patients. *Journal of neurosurgery*, *101*(3), 449-457.

MacGregor, L.J., Pulvermüller, F., van Casteren, M. & Shtyrov, Y. (2012). Ultra-rapid access to words in the brain. *Nature Communications,* 3, 7-11. doi:10.1038/ncomms1715

Mahon, B.Z. & Caramazza, A. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content *J. Physiol. Paris* 102, 59-70 (2008). doi:10.1016/j.jphysparis.2008.03.004

Mesulam, M. M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology,* 28, 597–613.

Milner, P. M. (1974). A model for visual shape recognition. *Psychological review*, *81*(6), 521.

Miozzo, M., Pulvermüller, F., & Hauk, O. (2014). Early parallel activation of semantics and phonology in picture naming: Evidence from a multiple linear regression MEG study. *Cerebral Cortex*, bhu137. doi: 10.1093/cercor/bhu137

Munding, D., Dubarry, A-S. & Alario, X-F. (Accepted). On the cortical dynamics of word production: a review of the MEG evidence. *Language, Cognition and Neuroscience.* DOI: 10.1080/23273798.2015.1071857

Näätänen, R., Paavilainen, P., Rinne, T. & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical Neurophysiology,* 118, 2544-2590. doi:10.1016/j.clinph.2007.04.026

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, *13*(11), 4700-4719.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, *31*(6), 785-806. doi:10.1016/0749-596X(92)90039-Z

Palm, G. (1980). On associative memory. *Biological cybernetics*, *36*(1), 19-31.

Palomar-García, M. Á., et al. (2015). Do bilinguals show neural differences with monolinguals when processing their native language? *Brain and language*, *142*, 36-44. doi:10.1016/j.bandl.2015.01.004

Perani, D., et al. (1998). The bilingual brain. *Brain*, *121*, 1841-1852. http://dx.doi.org/10.1093/brain/121.10.1841

Poulisse, N., & Bongaerts, T. (1994). First language use in second language production. *Applied linguistics*, *15*(1), 36-57. doi: 10.1093/applin/15.1.36

Price, C. J. (2012). A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage*, *62*(2), 816-847. doi:10.1016/j.neuroimage.2012.04.062

Pulvermuller, F. (1999). Words in the brain's language. *Behavioral and Brain Science,* 22, 253-336. DOI: http://dx.doi.org/10.1017/S0140525X9900182X

Pulvermuller F. (2002). *The neuroscience of language*. Cambridge University Press, Cambridge.

Pulvermuller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience* 6, 576-582. doi:10.1038/nrn1706

Pulvermuller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, 17, 458-470. doi:10.1016/j.tics.2013.06.004

Pulvermuller, F. & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience,* 11, 351–360. doi:10.1038/nrn2811

Pulvermuller, F., Garagnani, M., & Wennekers, T. (2014). Thinking in circuits: toward neurobiological explanation in cognitive neuroscience. *Biological cybernetics*, *108*(5), 573-593. 10.1007/s00422-014-0603-9

Pulvermuller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. *Brain and Language*, 110, 81–94. doi:10.1016/j.bandl.2008.12.001

Pulvermüller, F. et al. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci.USA* 103, 7865-7870. http://dx.doi.org/10.1073/pnas.0509989103

Roux, F. E., & Trémoulet, M. (2002). Organization of language areas in bilingual patients: a cortical stimulation study. *Journal of Neurosurgery*, *97*(4), 857-864.

Runnqvist, E., Strijkers, K., Sadat, J., & Costa, A. (2011). On the temporal and functional origin of L2 disadvantages in speech production: A critical review. *Frontiers in psychology*, *2*. http://dx.doi.org/10.3389/fpsyg.2011.00379

Runnqvist, E., Strijkers, K., Alario, F-X., & Costa, A. (2012). Cumulative semantic interference is blind to language: Implications for models of bilingual speech production. *Journal of Memory and Language,* 66, 850-869. doi:10.1016/j.jml.2012.02.007

Schomers, M. R., Kirilina, E., Weigand, A., Bajbouj, M., & Pulvermüller, F. (2014). Causal influence of articulatory motor cortex on comprehending single spoken words: TMS evidence. *Cerebral Cortex*, bhu274. doi: 10.1093/cercor/bhu274

Shadlen, M. N., & Newsome, W. T. (1994). Noise, neural codes and cortical organization. *Current opinion in neurobiology*, *4*(4), 569-579. doi:10.1016/0959-4388(94)90059-0

Siegel, M., Donner, T. H., & Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews Neuroscience*, *13*(2), 121-134. doi:10.1038/nrn3137

Simmons, W. K., et al. (2007). A common neural substrate for perceiving and knowing about color. *Neuropsychologia*, *45*(12), 2802-2810. doi:10.1016/j.neuropsychologia.2007.05.002

Singer, W. (1999). Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, *24*(1), 49-65. doi:10.1016/S0896-6273(00)80821-1

Singer, W. & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review Neuroscience,* 18, 555–586. DOI: 10.1146/annurev.ne.18.030195.003011

Singer, W. (2013). Cortical dynamics revisited. *Trends in cognitive sciences*, *17*(12), 616-626. doi:10.1016/j.tics.2013.09.006

Strijkers K., Baus C., Runnqvist E., Fitzpatrick I., Costa A. (2013). The temporal dynamics of first versus second language production. *Brain and Language*, 127, 6–11. doi:10.1016/j.bandl.2013.07.008

Strijkers, K., Bertrand, D., & Grainger, J. (2015). Seeing the Same Words Differently: The Time Course of Automaticity and Top–Down Intention in Reading. *Journal of Cognitive Neuroscience*, 8, 1542-1551. doi:10.1162/jocn_a_00797

Strijkers, K. & Costa, A. (2011). Riding the lexical speedway: A critical review on the time course of lexical selection in speech production. *Frontiers in Psychology*, 2, 356. doi: 10.3389/fpsyg.2011.00356

Strijkers, K., & Costa, A. (2016a). The cortical dynamics of speaking: Present shortcomings and future avenues. *Language, Cognition and Neuroscience*, 1-20. 10.1080/23273798.2015.1120878

Strijkers, K., & Costa, A. (2016b). On words and brains: linking psycholinguistics with neural dynamics in speech production. *Language, Cognition and Neuroscience*, 1-12. 10.1080/23273798.2016.1158845

Strijkers, K., Costa, A., & Pulvermuller, F. (2017). The cortical dynamics of speaking: Lexical and phonological knowledge simultaneously recruit the frontal and temporal cortex within 200 ms. https://doi.org/10.1016/j.neuroimage.2017.09.041

Strijkers, K., Costa, A., & Thierry, G. (2010). Tracking lexical access in speech production: electrophysio- logical correlates of word frequency and cognate effects. *Cerebral Cortex,* 20, 912–928. doi: 10.1093/cercor/bhp153

Strijkers, K., Holcomb, P., & Costa, A. (2011a). Conscious intention to speak facilitates lexical access during overt object naming. *Journal of Memory and Language*, *65,* 345–362. doi:10.1016/j.jml.2011.06.002

Tettamanti, M., et al. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of cognitive neuroscience*, *17*(2), 273-281. doi:10.1162/0898929053124965

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences*, *106*(11), 4567-4570. doi: 10.1073/pnas.0811155106

Thierry, G. & Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 12530–12535. doi: 10.1073/pnas.0609927104

Tsodyks, M. V., & Markram, H. (1997). The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences*, *94*(2), 719-723.

Van Essen, D. C., & Maunsell, J. H. (1983). Hierarchical organization and functional streams in the visual cortex. *Trends in neurosciences*, *6*, 370-375. doi:10.1016/0166-2236(83)90167-4

van Rossum, M. C., Turrigiano, G. G., & Nelson, S. B. (2002). Fast propagation of firing rates through layered networks of noisy neurons. *The Journal of neuroscience*, *22*(5), 1956-1966.

Varela, F. et al. (2001) The brainweb: phase synchronization and large-scale integration. *Nature Reviews Neuroscience,* 2, 229–239. doi:10.1038/35067550

Van Hell, J. G., & De Groot, A. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association.*Bilingualism: Language and Cognition*, *1*(03), 193-211. http://dx.doi.org/10.1017/S1366728998000352

Von der Malsburg, C. (1985). Nervous structures with dynamical links. *Ber. Bunsenges. Phys. Chem*, *89*(703-710), 1.

Wilson, S.M., Saygin, A.P., Sereno, M.I. & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience,* 7, 701-702. doi:10.1038/nn1263

Figure Captions

Figure 1. Schematic visualization of a sequential hierarchical spatiotemporal dynamic of word production. (A) Dominant spatiotemporal map of monolingual word production (*based on I&L, 2004*). Brain activity sequentially progresses from the left mid temporal gyrus (MTG) linked to lexico-semantic processes, towards the left superior temporal gyrus (STG) involved in lexical phonology, over the left inferior frontal gyrus (LIFG) associated with motor phonology and ending in the motor cortex (MC) for sending an articulation program. Numbers in the different processing hierarchies represent approximates in ms of when brain activity is thought to be maximal in that particular region. *(taken from Strijkers & Costa, 2016a).* **(B)** Adaptation of this model to bilingual word production. Brain activity follows the same sequential progression over hierarchically organized brain regions and linguistic components. L2 representations are plotted onto the brain template and follow similar anatomy-function color-coding as the L1 representations (with lighter or darker shaded colors). Activation time-course is expected to go

somewhat slower than in the monolingual version and the bigger areas for L2 in LIFG and the motor cortex depict the more extensive recruitment in those regions during L2 speech.

Figure 2. Schematic visualization of a binding-by-synchrony and Hebbian-like cell assembly network of word production. A widely distributed lexico-semantic network embedded in local-specific action (red) – perception (yellow) circuits and a widely distributed phonological-phonemic network embedded in specific action (blue) – perception (green) circuits form a word assembly which ignites as a whole within the first 200 ms of processing. After ignition, activity may remain active in the whole word assembly or reverberate in specific parts of the assembly to generate well-timed (sequential) spatiotemporal dynamics (e.g., reverberation in the lexico-semantic sensorimotor circuits for semantic integration and reverberation in the phonological-phonemic sensorimotor circuits for timed articulation) (*taken from Strijkers & Costa, 2016a*).

Figure 3. Bilingual adaptation of a (schematic) neural assembly based framework on words in the brain. The model follows the same structure (lexico-semantic and phonological-articulatory action-perception binding) and dynamic as the monolingual version, but with an assembly fitted to the brain template for each language (similar color-coding where lighter/darker shade differences depict the L1 versus L2 assembly). In this particular example, the map reflects concrete non-cognate words in balanced bilinguals, in which case local language differences are most apparent in STG, LIFG and motor cortex associated with phonological-articulatory properties of a word.

Figure 4. Specific neural assembly maps related to language membership, cognate-status and lexico-semantics. (A) Schematic map of L1 and L2 assemblies (for simplicity colored in black and gray only) where consistent and repeated variability in acoustic-articulatory properties between L1 and L2 words form

supra-assemblies coding language membership in STG, LIFG and motor cortex (for visibility the L1 and L2 "nodes" are somewhat clustered, but in reality these will concern distributed patterns in those regions). **(B)** Schematic map of L1 and L2 assemblies coding for the cognate-status of words, where for non-cognates there is little overlap in STG, LIFG and motor cortex, for non-identical cognates there is considerable overlap, and for identical cognates there is almost complete overlap (for simplicity the assemblies were "stripped" from their lexico-semantic part). **(C)** Schematic map of L1 and L2 assemblies reflecting potential lexico-semantic variations between words. For example, in contrast to many concrete nouns, certain emotion, taboo or abstract words could display distributional differences between L1 and L2 assemblies in brain regions such as the orbito-frontal cortex, the hippocampus and subcortical structures (for simplicity the assemblies were "stripped" from their phonological-articulatory part).