The Government of the Russian federation

Federal State Autonomous Educational Institution of Higher Professional Education

National Research University – Higher School of Economics

Faculty of Social Sciences, School of Psychology,
Master's program
"Cognitive sciences and technologies: from neuron to cognition"

Coursework

«Biologically Inspired Neurocomputational Model of Semantic Representations»

Student
Kuptsova, Anastasia
Last name, First name Middle name

_____
Signature

Scientific supervisor
Professor, PhD
Position, Academic degree
Gutkin, B.
Last, F. M./O.

Advisor
Professor, PhD
Position, Academic degree
Shtyrov, Y.
Last, F. M./O.

Moscow, 2019

Table of Contents

Chapter 1. Introduction

Semantic knowledge is the knowledge about meaning of concepts and words (Quillan, 1966; Kiefer & Pulvermüller, 2012). We use it constantly to behave properly in different situations, to express our thoughts, to understand others and so on. People with semantic memory impairments have difficulties with word comprehension and retrieval (Hodges & Patterson, 2007; Montembeault, Brambati, Gorno-Tempini, & Migliaccio, 2018). They can forget how to use ordinary things, for example, an umbrella (Murre, Graham, & Hodges, 2001); or they cannot understand and explain the source of pain that they feel (Montembeault et al., 2018). Therefore, semantic knowledge constitutes the core units of our thoughts and behavior. And maybe revealing semantic system organization would allow us to make a step toward solving the eternal mystery of the human cognition.

Early theories assumed that the semantic system consists of symbolic representations of real-world entities in the brain (e.g. each semantic concept is just a node in the network (Quillian, 1969; Collins & Loftus, 1975); or each concept could be represented by the features list (Smith, Shoben, & Rips, 1974)), and to use our semantic memory we should manipulate these symbols. The problem with these approaches was in the question how our real-world experience through which we acquire concepts can be transduced into symbols, which are totally distinct from this sensory-motor experience (Harnad, 1990) — in other words, how our brain can acquire and store amodal representations of concepts (Barsalou, 1999). In 90's the first evidence, as well as theories utilizing it, of concepts being grounded in sensory and motor modalities appeared (Barsalou, 1999; Pulvermüller, 2001; Damasio, Grabowski, Tranel, Hichwa, & Damasio, 1996). However, different research groups debated about different degrees of grounding. Some theories posit that sensory-motor perception plays only a secondary role in the semantic representations (Mahon & Caramazza, 2008; Rogers et al., 2004; Patterson, Nestor, & Rogers, 2007) and that it is important to have one higher-order brain area where all concepts will be stored in an amodal way; others posit that when we process the concept, our brain almost simulates the perceptual experience of the interaction with this concept (Gallese &

Lakoff, 2005); in between, there are hybrid theories where both the necessary grounding on the perceptual experience and higher-order zones to mediate the experience from different modalities are suggested (Kiefer & Pulvermüller, 2012; Garagnani & Pulvermüller, 2016; Binder & Desai, 2011). Although there exist criticisms of all theories, current empirical evidence speaks in favor of hybrid ones (Kiefer & Pulvermüller, 2012; Meteyard, Cuadrado, Bahrami, & Vigliocco, 2012; Binder, Desai, Graves, & Conant, 2009).

Different approaches to model semantic system were proposed recently. However, the present-day situation in this filed is the following: two extreme approaches to the semantic representation modeling and the niche for further research in between. One extreme is presented by the data-driven but purely statistical models, in which correlation between the real-word inputs and cortex activity is computed (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Anderson et al., 2016). Another extreme is biologically accurate neurocomputational models, which get as input only abstract synthetic data (Chen, Lambon Ralph, & Rogers, 2016; Garagnani & Pulvermüller, 2016; Tomasello, Garagnani, Wennekers, & Pulvermuller, 2018). The bridge between real data and bio-inspired architecture of the model is currently missing.

The plausible approach to build this bridge is to improve and extend the most biologically accurate models to shift them into the real data domain. For instance, static and context-independent input patterns are usually used in such models (Chen et al., 2016; Garagnani & Pulvermüller, 2016; Tomasello et al., 2018). However, in the real world, we perceive words in dynamically changing contexts. The possibility of a dynamic and context-dependent input patterns would allow for better correspondence between the model and the real semantic system. Moreover, very limited amount of cortex areas in the left hemisphere are usually modeled (Chen et al., 2016; Garagnani & Pulvermüller, 2016; Tomasello et al., 2018). Therefore, adding more cortex areas in both hemispheres and subcortical structures which also contribute to semantic memory can be beneficial.

However, before shifting the most biologically accurate model to the real data domain, important qualitative results should be checked for this model — to understand whether this "best" model is worthwhile to work on or if there are critical problems in it. This is exactly what different research groups are working on. For example, models' predictions are checked in the modelled congenitally blind individuals (Tomasello, Wennekers, Garagnani, & Pulvermüller, 2019; Chen et al., 2017) or in the modelled patients with different lesions and diseases (Ueno, Saito, Rogers, & Lambon Ralph, 2011; Chen et al., 2017).

The goal of this paper is in line with the qualitative testing idea discussed above. I will replicate the most biologically precise model of semantic representation which exists to date (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018) and will add the mechanisms of semantic dementia, which is the core semantic disease, to it (Hodges, & Patterson, 2007; Montembeault et al., 2018).

This paper is organised in the following way. In the literature review, theories of semantic representation and neurocomputational attempts to model semantic system are discussed; and also semantic dementia symptoms and neuronal changes are presented. In the proposal part, I describe the most biologically precise model and my idea regarding how to add semantic dementia mechanism to this model.

Chapter 2. Literature Review

2.1 Theories of Semantic Representations

During the second part of the 20th-century, scientists started to try to reveal the mechanisms and principles of the human cognition. And at the same time scientists were trying to use these “cognitive” mechanisms in the building of artificial intelligence, as they believed they could build it very soon and make it very biologically grounded. However, for this reason, early cognitive theories were attempted to simplify the explanations of the cognitive mechanisms to the point where existing then computers could simulate them (Barsalou, 1999; Meteyard et al., 2012). The symbolic approach to semantic representation was one of such early cognitive theories.

The symbolic approach states that the perceptual sensory-motor experience which we get while interacting with the real world entities or events is transduced into the symbolic representations and then these representations are used in different aspects of cognition. One theory proposed in this vein was the semantic network model (Quillian, 1969; Collins, & Loftus, 1975). Quillian (1969) built a computational program for mechanical extraction of the meaning from the text. He argued that theory which underlies this program could be referred to semantic processing by human as well. Quillian (1969) and Collins and Loftus (1975) presented the semantic memory model as the network with distinct nodes, which symbolize the concepts, and links between the related nodes (concepts). Different types of links in this network reflect the different relationship between concept, for instance, superordinate or subordinate category. To illustrate the idea, the sentence “apple is red” would be processed in their network in the following way: two nodes "apple" and "red" are connected with the link “is”, which shows that color attribute belongs to the concept. Another example of the model, which was influenced by the symbolic approach, is the features list model (Smith, 1974). The author suggested that the meaning of the word is represented by the list of features. This list is flexible, as some features are more common (probable) for the particular concepts than others.

For example, the concept “apple” could be defined by the following features — round, sour-sweet, grows on a tree, green and so on — here green is a less common feature than round.

Another branch of the early semantic theories originates from the connectionist approach (McClelland, Rumelhart, & PDP Research Group, 1986; Smolensky, 1988; for review see Kiefer & Pulvermüller, 2012). According to this view, semantic system could be defined as the network with neuron-like units, some of which are connected via links. These links have weights, which are dynamically changing under the input patterns and the learning rule for the network. In this approach, concepts could be represented as the pattern of activity propagation through the network units. This idea is similar to the modern neural networks approach in computer science or to bio-inspired neuronal networks (for example, see Garagnani & Pulvermüller, 2016; Chen et al., 2017).

In both the symbolic and the early connectionist approaches, semantic representations were amodal — distinct from the sensory-motor experience and from its circuits in the brain (Barsalou, 1999). However, this amodal view raises a very important and reasonable question: what are the neural mechanisms of extraction of the amodal representations from the real world entities and shouldn’t meaningfulness imply at least some grounding on the interaction with the real world? This question is known as the “symbol grounding problem” (Harnad, 1990). In the ‘90s the first pivotal works about semantic grounding on the sensory-motor system appeared (Barsalou, 1999; Pulvermüller, 2001; Damasio et al., 1996). It was shown (in the patients with focal cortical lesions or with the help of the neuroimaging techniques) that comprehension and retrieval of the different categories of concepts (e.g. animals vs tools) correlate with different cortical sites (e.g. occipital and inferior temporal vs premotor, accordingly) (for review see Barsalou, 1999; Pulvermüller, 2001). It was assumed that the network of the strongly connected neurons in the sensory and motor areas, which was built from the sensory-motor associations during concept acquisition, could create a basis for semantic representations. To illustrate the idea, the concept “apple” could be represented by the distributed sensory-motor

associations like the visual, gustatory, and olfactory perception of an apple, auditory perception of its name “apple”, motor knowledge of how to grasp an apple and so on. Barsalou (1999) proposed that, while the perception phase, primary sensory-motor areas send the ascending perceptual patterns to the “associated areas”, which in turn, during comprehension phase, activate other sensory-motor areas to complete the semantic representation. Damasio and colleagues (1996) suggested that "higher order association cortices" are located near the primary motor and sensory areas and mediate the word processing (during word retrieval or comprehension). These neuroimaging findings and theoretical proposals of grounding leave very little doubt that the conceptual system is somehow connected with the sensory-motor one — therefore the conceptual system is embodied. However, it raises one of the main questions for the modern research of semantics: to which extent is the conceptual system embodied?

One extreme answer to this question proposes that sensory-motor activation plays only the secondary or complementary role with respect to the amodal representation of the concept. Mahon and Caramazza (2008) argue that today almost no one suggests that the sensory-motor system is absolutely distinct from the conceptual one. However, the dynamic of activation between and within perceptual and conceptual systems should be studied meticulously. Hypothetically, the visual perception of an apple could (1) directly activate fully distributed sensory-motor representation of the apple concept; or (2) firstly activate distributed sensory-motor representation, which in turn activates amodal representation of the apple concept; or (3) firstly activate amodal representation, which in turn activates distributed sensory-motor representation of the apple concept. This third dynamic is assumed by Mahon and Caramazza (2008) in their "grounding by interaction" hypothesis. They argue that after the creation of the amodal representation, sensory-motor activation completes the full concept representation. A similar idea is proposed by Rogers and colleagues as a “hub-and-spoke” model (Rogers et al., 2004; Patterson et al., 2007; Lambon Ralph, Jefferies, Patterson, K., & Rogers, 2017). They suggest that activation in the sensory, motor and language areas only — spokes — is not enough to create the

semantic system, as problems with the concept generalization should appear. For example, apple and kiwi have different name, shape, color, taste and so on, however, we somehow generalize them into the one fruit category. To integrate all attributes of the concept, the higher-order zone in the anterior temporal lobe (ATL) — hub — is needed. The idea of the single amodal semantic hub in the ATL, deterioration of which correlates with semantic comprehension problems in patients with semantic dementia, is the core part of this theory. To illustrate the hub-and-spoke model, visual perceptual information of an apple initially flows into the amodal hub in the ATL, where "intermediate representation" of the apple appears and associations with all other perceptual modalities are stored (e.g. how to grasp an apple).

Another extreme answer to the question about the extent of embodiment includes "fully distributed" or "strong embodiment" theories (this name was borrowed from Binder, & Desai, 2011; Meteyard et al., 2012). Gallese and Lakoff (2005) refer to the neuroimaging findings, which show that imaging and actioning share the same neuronal site. Then they continue that imaging some action and knowing about this action share the same neuronal site, as it is hard to imagine that if we can not imagine some action, we can have knowledge about it. So they conclude that actioning and having the knowledge about this action share the same neuronal site. Furthermore, similar logic (and the appropriate empirical findings) could be used to prove that objects representations ground in the sensory-motor system. Therefore the authors infer that the structure of the sensory-motor system itself can be used for semantic representations and representation of the concept is the simulation of the perceptual experience of this concept.

In between of these extremes, there exist hybrid theories which assume both the essential grounding of the concept on the sensory-motor activation and the need of the higher order convergence zones to connect and mediate this embodied experience. Pulvermüller and colleagues (Kiefer, & Pulvermüller, 2012; Garagnani & Pulvermüller, 2016) suggest that concept knowledge consists of the (1) modality-specific representations, which were acquired from the perceptual experience and are coded in the distributed sensory-motor areas of the cortex, and (2) "connector hubs",

which help with mediation and connection between modalities. They argue that this hypothesis follows directly from the perceptual origin of the semantic knowledge and the neurobiological properties of the brain (Pulvermüller, 2001; Garagnani & Pulvermüller, 2016). When we acquire the concept, the correlation occurs between different perceptual modalities (e.g. mom teaches her child “This is an apple ”, so in this case the auditory and visual perception emerge at the same time). This perceptual co-occurrence could be used by the Hebbian learning mechanism (“fire together — wire together”) to establish robust links between different perceptual modalities (Artola & Singer, 1993). However, the patterns of the cortex connectivity imply that anatomically there should be mediatory areas between the primary ones, therefore “connector hubs” are needed. In the “embodied abstraction” hypothesis Binder and Desai (2011) propose that higher-order “convergence zones” are located in temporal and inferior parietal areas. They assume that conceptual representations have different hierarchical levels (a general level and a more detailed one), which depend on context, concept familiarity and task demand. During the general level of representations more abstract higher-order cortices are activated, while for the detailed representations low perceptual mechanisms are needed.

As I mentioned previously, the aim for the present-day researchers is to understand to which extent the conceptual system is embodied. Further, I would like to discuss the most important critiques and bottlenecks of different theories of the embodied semantics which were presented above.

The main problem of secondary embodied theories (this name was borrowed from Meteyard et al., 2012), like grounding by interaction theory (Mahon, & Caramazza,2008) or hub-and-spoke model (Rogers et al., 2004; Patterson et al., 2007; Lambon Ralph et al., 2017), is that they cannot fully explain the current data about the correlation between the activity in the different cortical sites and the comprehension of the specific categories of concepts. For instance, patients with lesions in the motor and premotor cortex could suffer from the impaired comprehension of the action-related words compared with the object-related ones, while patients with lesions in the visual cortex have problems with comprehension of

the object-related words (for review see Kiefer, & Pulvermüller, 2012; Meteyard et al., 2012). Furthermore, it was shown that transcranial magnetic stimulation of the hand motor areas could enhance comprehension of the hand-related words, while stimulation of the leg areas could do the similar thing with the leg-related words (Pulvermüller, Hauk, Nikulin, & Ilmoniemi, 2005). Let us assume the following situation: the patient has the lesion in the primary motor area. She is presented with the picture of the scissors and asked to name it. According to these secondary embodied theories, the visual perception of the scissors will propagate to the amodal hub, where the "intermediate representation" of the scissors appears and associations with other perceptual representations are stored (among which there is an association between the view of the scissors and its name "scissors"). Therefore, it seems that in this paradigm there should not appear any problems with the name "scissors" retrieval, as the lesion in the motor area doesn't influence the "intermediate representation" in the amodal hub and the association between the visual perception of the scissors and its name. But this contradicts empirical results.

The central flaw of the fully distributed theories (Gallese, & Lakoff, 2005) is that they cannot explain the current data about the existence of the cortical areas crucial for semantic processing, which are distinct from the sensory-motor system (therefore, from the perceptual experience). For instance, patients with semantic dementia have progressive ATL atrophy, which is correlated with the semantic knowledge deterioration, however, the ATL is not located in the sensory-motor system (Hodges, & Patterson, 2007; Montembeault et al., 2018; Spinelli et al., 2017). Moreover, the review of the 120 fMRI studies shows that there exist multiple cortical areas, which are important for general semantic processing, for instance, in the lateral and ventral temporal areas and in the inferior parietal area (Binder et al., 2009).

The core critique toward the hybrid theories (Kiefer, & Pulvermüller, 2012; Garagnani & Pulvermüller, 2016; Binder, & Desai, 2011) is that the activation of the sensory-motor areas during the semantic processing does not prove that the same neuronal mechanism as in perceptual system is used in the semantic system, maybe these areas simply overlap (Mahon, & Caramazza, 2008; Caramazza, Anzellotti,

Strnad, & Lingnau, 2014). However, the following argument could be used against this critique: as it was proposed by Pulvermüller and colleagues (Pulvermüller, 2001; Garagnani & Pulvermüller, 2016), for concepts acquisition our brain uses correlation learning mechanism during co-occurrence of the activation in the different perceptual modalities. If concepts originate from the correlation between different perceptual modalities, how can our brain totally isolate concepts from the perceptual system after acquisition via perceptual co-occurrence is finished?

2.2 Approaches to Semantic Representations Modeling

As I mentioned earlier, the present-day approaches to model semantic representations consist of the two extremes and the niche for further research in between. One extreme is the purely statistical models, which correlate the meaning of the real words with the brain activity across the whole cortex (Huth et al., 2016; Anderson et al., 2016). Another extreme is the bio-inspired neurocomputational models, which are presented by the limited number of the cortex areas and get only abstract synthetic data as input (Chen et al., 2017; Garagnani & Pulvermüller, 2016; Tomasello et al., 2018).

Huth et al. (2016) tried to create the map of the semantic distribution across the cortex. During the experiment, seven participants were listening to a story for more than two hours in the fMRI scanner. Each word of the story was projected into the 985-feature space, which consists of the 985 popular English words. The projection was based on the similarity of the context between the word in the story and the each of the 985 popular English words (e.g. “giraffe” and “zebra” are often used in the similar context when “giraffe” and “socket” are not). BOLD signal in each voxel of the cortex was regressed on the 985 features and controls, which helps to catch only the semantic component of the words and avoid others, for example, phonetic one. To reveal which semantics activate each voxel in a more convenient way — therefore to create the semantic map — authors further performed principal component analysis (PCA) and clustered all the words. They got 12 categories, which they labeled manually in the following way: tactile, visual, numeric, locational, abstract, temporal, professional, violent, communal, mental, emotional. They created a semantic map of

the cortex in accordance with these categories and showed that this map is more or less consistent across different individuals. A very similar approach to create a semantic map was proposed by Anderson et al. (2016). They also used fMRI recording (in this case, during the sentence reading) and applied a purely statistical algorithm (multiple regression) to predict the activity to particular words in the cortex (semantic representations). However, they used predefined in the previous study "semantic features" (Binder et al., 2016), which were distributed across the different categories: visual, auditory, somatosensory, gustatory, somatosensory, motor, attentional, event, evaluation, cognitive, emotional, drive, spatial.

On the other extreme Chen et al. (2016) suggested that the pattern of semantic representations could be explained by the combination of three factors: the sensory-motor experience, brain connectivity, and associative learning. They implemented these ideas in a model, which is a biologically constrained deep recurrent neuronal network. It should be noted, that the idea of the hub-and-spoke model underlies this architecture (Rogers et al., 2004; Patterson et al., 2007). The model consists of three modality-specific regions — superior temporal gyrus for auditory representation (auditory region); parietal lobes, posterior middle temporal gyrus and medial posterior fusiform gyrus for functions and praxis representation (motor region); lateral posterior fusiform gyrus for visual representation (visual region) — and one amodal hub in the anterior temporal lobe (ATL). The choice of these areas was based on the previous studies, which revealed the most important brain regions for semantic processing; connectivity between all these areas was assessed via DTI. Further, the authors synthesized two different semantics — animals and tools. It was suggested that animal comprehension relies mostly on the visual experience, therefore, the model was taught to associate activation in the visual and auditory areas to create semantic representations of animals. On the contrary, for tools, understanding of the motor experience is more important than the visual one, therefore, the model was taught to associate activation in the function-praxis and auditory areas and reduced activation in the visual areas to create semantic representations of tools. After the learning phase, the model showed higher activation in the visual regions for animals

input and higher activation in the function-praxis regions for tools input. This result is consistent with the empirical data, in which correlation between the different cortical sites activation and processing of the specific semantic categories is shown.

A similar but significantly more biologically accurate model was proposed by Pulvermuller and colleagues (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018). They also suggested that sensory-motor co-experience during concept acquisition, brain connectivity, and associative learning mechanism are fundamental principles which can explain the empirical data about semantic representations. However, authors did not assume a separate network consisting of predefined locations for semantic representations, as was done by Chen et al. (2016). On the contrary, they tried to simulate different cortical regions and to show how the semantic system could emerge on the basis of these regions and fundamental principles of the brain anatomy and functioning. The model consists of four modality-specific zones, with each zone containing three cortical areas — from primary to "central" ones: auditory areas (primary auditory cortex, auditory belt and auditory parabelt), articulatory areas (inferior primary motor cortex, inferior premotor cortex and inferior prefrontal cortex), visual areas (primary visual cortex, temporo-occipital cortex and anterior temporal cortex) and motor areas (lateral primary motor cortex, lateral premotor cortex and dorsolateral prefrontal cortex). Accurate neuroanatomical connectivity between these areas is one of the core features of this model and is a result of years-long research. Another important distinction of this model is biologically constrained fine architecture of the neuronal network and biologically constrained learning algorithm. Each area consists of the two layers of neurons — excitatory and inhibitory ones. Excitatory neurons are graded-response neurons (Garagnani & Pulvermüller, 2016) in the early model iteration and spiking neurons in the latest one (Tomasello et al., 2018); inhibitory neurons are graded-response neurons. Global and local inhibition mechanisms are implemented. Connections between and within areas are sparse, topographically constrained and random (with decreasing probability, which is depended on the distance between two neurons). Notably, the authors use Hebbian learning mechanism ("fire together — wire

together") — an associative learning mechanism which is assumed to be used in the brain. I will describe this model (and the Hebbian learning mechanism as well) in more detail in the *Proposal* part. All neurobiological features which I discussed above significantly distinguish this model from the model proposed by Chen et al. (2016), where a more mechanical recurrent neural network with backpropagation as the learning mechanism is used. The problem of such modeling (Chen et al., 2016) is that this architecture and learning mechanism do not try to replicate the real properties of the brain. Therefore, such model can only fit the empirical data but cannot provide the explanation of the processes in which this data appears.

Authors (Garagnani & Pulvermüller, 2016) synthesized two different semantics — object-related and action-related words. The model was taught to associate co-activation in the primary visual, primary auditory and primary articulatory (inferior primary motor) areas to create semantic representations of the object-related words. To acquire action-related words, the model was taught to associate co-activation in the primary motor (lateral primary motor cortex), primary auditory and primary articulatory (inferior primary motor) areas. After the learning phase, word recognition was assessed by presenting only the auditory pattern to the primary auditory area. It was shown that auditory pattern from object-related words propagates until the primary articulatory and visual areas when almost no activity is registered in the primary motor area. The propagation pattern is opposite for the motor-related words — primary articulatory and motor areas. This result means that the model learns to discriminate between different categories of semantics on basis of sensory-motor representations, which is consistent with the empirical data. Notably, the most "central" areas in visual and motor zones — anterior temporal cortex and dorsolateral prefrontal cortex — do not distinguish between two categories of semantics, therefore, they are category-general. This is consistent with the empirical findings that there exist several multimodal "connector hubs" (Binder & Desai, 2011; Kiefer & Pulvermüller, 2012).

2.3 Semantic Dementia

Semantic dementia (SD) is a progressive neurological disease, which is characterized by semantic knowledge deterioration. The degree of this deterioration is correlated with the severity of the anterior temporal lobe (ATL) atrophy (Hodges & Patterson, 2007; Montembeault et al., 2018; Spinelli et al., 2017). Even more, according to some computational models of SD (Ueno et al., 2011; Chen et al., 2017) and TMS studies (Pobric, Jefferies, & Lambon Ralph, 2007), the ATL atrophy is considered as a cause of the problems with the semantic memory. Therefore, SD is a core disease to explore how semantic system works and what neuronal mechanisms underlie it.

The main symptoms of SD are anomia and the difficulty with word comprehension (for review see Hodges & Patterson, 2007; Montembeault et al., 2018). Anomia is difficulty with the retrieval and usage of the right words, for instance, when the SD patient tries to describe something or to tell the story. When an SD patient has difficulties with word comprehension, she misunderstands some words or the whole speech. Moreover, such people have difficulties with understanding when they have to interact with concepts in a non-verbal from. For example, the SD patient can forget how to use an umbrella (Murre et al. 2001), or she can draw a duck as an animal with four legs (Patterson & Erzinçlioğlu, 2008). These cognitive impairments are progressive — they become worse each year. Notably, SD patients have relatively preserved phonology and grammar, as well as episodic and autobiographical memory, at least until the late stage of the disease. However, during the middle and late stages the overall behavioral changes appear — apathy, irritability, lack of empathy and so on (for review see Hodges & Patterson, 2007; Montembeault et al., 2018). The median of the survival after the diagnosis with SD is 12 years (Hodges et al., 2009).

Hodges and colleagues studied and carefully documented many SD cases (Hodges, Patterson, Oxbury, & Funnell, 1992; Hodges, Graham, & Patterson, 1995; Hodges, &Patterson, 2007; Hodges et al., 2009). Therefore, from their works we

know about the typical errors which SD patients make (Hodges et al., 1995). SD patients could name the object of interest with the other object name, when the other object is in the same category as the object of interest — for example, “zebra” for “giraffe”. Or, they could name the object of interest with the superordinate category name — for example, “animal” for “zebra”. Or, they could describe the object of interest instead of naming it — for example, “the horse in the desert” for “camel”. Or, they could name the semantic association instead of the word itself (“tree” for “forest”). Or, they could name the object of interest with the other object name when the other object is in the other category than the object of interest (“apple” for “bird”).

Furthermore, there exist a highly robust dynamical pattern of the semantic deterioration — progressive loss, which starts from the degradation of specific semantic details and spreads to more general ones. Firstly, differences between close specific categories are deteriorated — for example, “zebra” for “giraffe”. Then the superordinate category could be used instead of the target one — for example, “animal” for “giraffe”. Finally, the patient just asks "what does giraffe mean?" (Hodges et al., 1995; Hodges & Patterson, 2007). Moreover, less frequent or familiar semantics, as well as more atypical semantics, are deteriorated earlier during the disease course (Hodges & Patterson, 2007; Montembeault et al., 2018; Rogers, Patterson, Jefferies, & Lambon Ralph, 2015).

Other patterns of the semantic deterioration need to be studied better, as there is no consensus about them across different works. For instance, in some studies the reverse concreteness effect is found (Yi, Moore & Grossman, 2007; Joubert et al., 2017) — the concrete concepts are deteriorated earlier during the disease course than the abstract ones — while other studies find the opposite effect (Hoffman & Lambon Ralph, 2011). Furthermore, although a lot of scientists suggest that the semantic deterioration during SD is category-general (all categories are deteriorated almost in the same manner and at the same disease stage) (Patterson et al., 2007; Lambon Ralph, Lowe, & Rogers, 2007), Yi and colleagues (2007) have found that nouns

(object-related words) are relatively more preserved than action verbs during the disease course.

Neuronal anatomical and functional changes during SD can be described by a bilateral and asymmetrical pattern (Hodges & Patterson, 2007; Montembeault et al., 2018) of the progressive atrophy (Hodges & Patterson, 2007; Montembeault et al., 2018; Spinelli et al., 2017) and hypometabolism (Montembeault et al., 2018) in the white and gray matter. Gray matter changes are detected in the anterior temporal lobes — superior, middle, inferior temporal gyri, fusiform gyrus, tempral pole, parahyppocampal gyrus — and progress into the basal ganglia and the medial orbitofrontal cortex during the disease course (for review see Brambati et al., 2015; Spinelli et al., 2017). White matter changes are detected in the regions which are connected with or underlie the temporal lobes: left inferior fronto-occipital fasciculus; uncinate fasciculus and inferior longitudinal fasciculus bilaterally (for review see Brambati et al., 2015; Spinelli et al., 2017). The maximal (and the most robust among the SD patients) atrophy is detected in the anterior temporal pole and ventral parts of the ATL (Lambon Ralph et al., 2017).

The degradation is often bilateral and asymmetrical, and usually more severe in the left ATL (Hodges et al., 2009). However, asymmetrically more severe degradation in the right ATL occurs as well. There exists a consensus among studies that atrophy in the left ALT mostly leads to the anomia and difficulties with words comprehension, while atrophy in the right ALT results in problems with person recognition, other social interactions and behavior changes (Hodges & Patterson, 2007; Lambon Ralph et al., 2017; Montembeault et al., 2018).

Chapter 3. Proposal

The model which was proposed by Pulvermüller and colleagues (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018) is the most biologically accurate model of the semantic system existing today (for review see the *Approaches to modeling of semantic representation* in the *Literature review* section). I will use an early iteration of this model (Garagnani & Pulvermüller, 2016)[1]. My idea is to check model's qualitative predictions by adding a semantic dementia mechanism there.

It should be noted that there were several attempts to model semantic dementia previously (Ueno et al., 2011; Chen et al., 2017). However, these attempts were made in the hub-and-spoke paradigm with predefined areas for semantic representation, and these models were less biologically constrained (e.g. one of them uses an artificial recurrent neural network with backpropagation learning mechanism (Chen et al., 2017)). Thus, it appears that the previous models, even if fitting the data well, do not explain the biological mechanisms of semantic representations. Moreover, to the best of my knowledge, to date, there was no attempt to model the fine consequences of semantic dementia — progressive loss of semantic knowledge which starts from deterioration of specific semantic details and spreads to more general ones. Therefore, the current research could contribute to the understanding of the semantic system in the brain.

In the first part of this proposal, the baseline model — which is the synthesis of the model used by Pulvermüller and colleagues (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018) and my additions (initially, without semantic dementia mechanism) — is described in detail. I will reproduce this baseline model and check its results before transferring to the SD modeling. In the second part, the technical idea of the SD modeling and its expected results are described.

[1] Authors of this model kindly shared their source code with me, therefore, I could replicate their model in the detail

3.1 Description of the Baseline Model

3.1.1 Macro-structure. The original model (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018) consists of four modality-specific zones, with each zone containing three cortical areas — from primary to more "central" ones (see **Fig. 1**). The auditory zone comprises the superior and lateral auditory areas — primary auditory area (A1), auditory belt (AB) and auditory parabelt (PB). The articulatory zone comprises the inferior frontal areas — inferior primary motor area ($M1_i$), inferior premotor area (PM_i) and inferior prefrontal area (PF_i). The visual zone comprises the inferior temporo-occipital areas — primary visual area (V1), temporo-occipital area (TO) and anterior temporal area (AT). The motor zone comprises the superior-lateral frontal areas — lateral primary motor area ($M1_L$), lateral premotor area (PM_L) and dorsolateral prefrontal cortex (PF_L). In early works, only six areas of the traditional language cortex were modeled — auditory zone and articulatory zone (Garagnani, Wennekers, & Pulvermüller, 2007). This model was extended by six areas from the visual and motor zones, which allows grounding concept acquisition and representation on the sensory-motor experience — to build the basis for the semantic system (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018). I will use the four-zones model, as a baseline model.

Three types of cortico-cortical connections are suggested in the latest iteration of the model (Tomasello et al., 2018). The first type is connections between two adjacent areas (depicted by black arrows in the **Fig. 1**): A1-AB and AB-PB in the auditory zone; $M1_i$-PM_i and PM_i-PF_i in the articulatory zone; V1-TO and TO-AT in the visual zone; $M1_L$-PM_L and PM_L-PF_L in the motor zone; PB-AT between auditory and visual zones; PF_i-PF_L between the articulatory and motor zones. The second type is long-distance cortico-cortical connections (depicted by purple arrows in the **Fig. 1**): PF_i-AT between articulatory and visual zones; PB-PF_L between auditory and motor areas; PB-PF_i between auditory and articulatory zones; AT-PF_L between visual and motor areas. The third type is "jumping links" which connect the "second neighbor" areas (depicted by blue arrows in the **Fig. 1**): A1-PB in the auditory zone; AB-PF_i and PB-PM_i between the auditory and articulatory zones; PF_i-$M1_i$ in the

articulatory zone; V1-AT in the visual zone; TO-PF_L and AT-PM_L between the visual and motor zones; PF_L-$M1_L$ in the motor zone. The decision about including these connections is based on the decades-long research on the neuroanatomical structure of the cortex (for review see Tomasello et al., 2018). In the baseline model, I will use only two types of connections — connections between two adjacent areas and long-distance cortico-cortical connections, as it was done in the early iteration of this model (Garagnani & Pulvermüller, 2016)[2]. It was shown that early and latest iterations of the model give qualitatively similar results, although, the latest one is more biologically realistic. The choice of the early iteration of the model is done for the initial simplification of the baseline model.

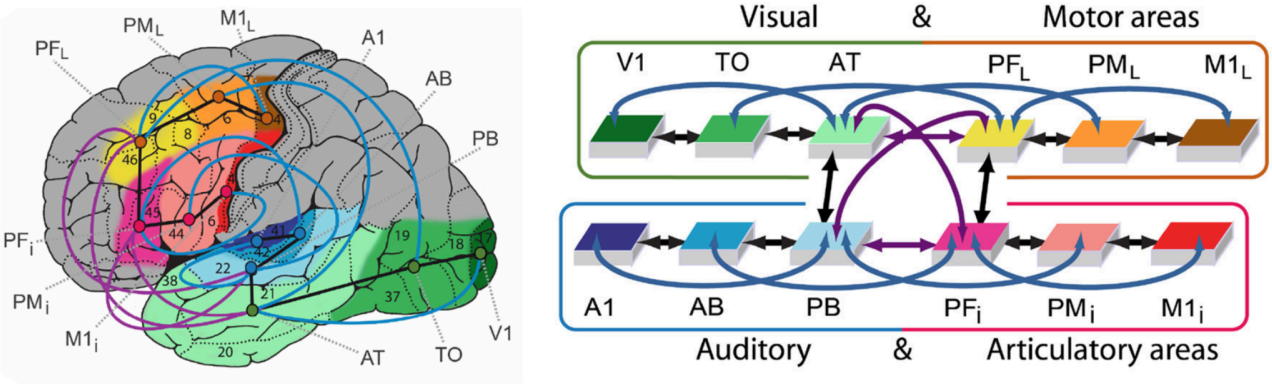

Figure 1. Four cortex zones (auditory, articulatory, visual and motor) with 3 areas within each zone and connectivity pattern (black arrows for connections between two adjacent areas; purple arrows for long-distance cortico-cortical connections; blue arrows for "jumping links"). Reprinted from "A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity", by Tomasello, R., Garagnani, M., Wennekers, T., & Pulvermuller, F, 2018, Frontiers in computational neuroscience, 12, 88. Copyright © 2018 Tomasello, Garagnani, Wennekers and Pulvermüller.

[2] Please note that compared with the early iteration, connections PB-AT and PF_i-PF_L will be added

3.1.2 Micro-structure. Each area consists of two neuronal layers — excitatory (e-cell) and inhibitory (i-cell) ones — with 625 (25x25) cells in one layer (see **Fig. 2**). Each i-cell corresponds exactly to one e-cell, a combination of e-cell and i-cell reflects approximately one cortical column — pyramidal excitatory neurons and inhibitory interneurons. Excitatory neurons are modeled as graded-response neurons (Garagnani & Pulvermüller, 2016) in the early model iteration and as spiking neurons in the latest one (Tomasello et al., 2018); inhibitory neurons are modeled as graded-response neurons. In the baseline model, I will use only graded-response neurons for the initial simplification.

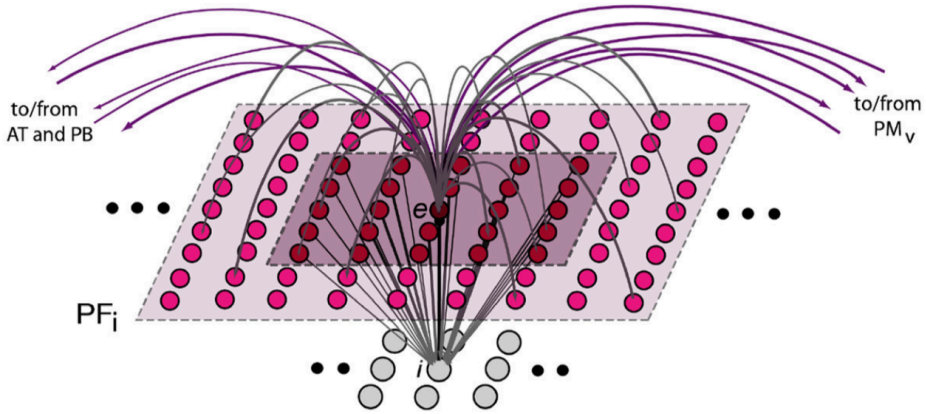

Figure 2. Micro-structure of the model, as an example from the PF_i area. Pink area reflects the e-cells layer with the e-cell of interest in the middle of the patch, grey units are i-cells with the i-cell in the middle, which correspond to e-cell of interest. Reprinted from "A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity", by Tomasello, R., Garagnani, M., Wennekers, T., & Pulvermuller, F, 2018, Frontiers in computational neuroscience, 12, 88. Copyright © 2018 Tomasello, Garagnani, Wennekers and Pulvermüller.

To understand the fine connectivity within and between areas, let us consider one of the e-cells (see **Fig. 2**). Firstly, the interaction of this e-cell with other e-cells will be explained. E-cell can send its projections in the 19x19 e-cells patch in the

same area, the e-cell under consideration being the central one in this patch. It also can send its projections in a topographically similar 19x19 e-cells patch in another area (with which the current area is connected through adjacent or long-distance cortico-cortical connection, as discussed in the section Macro-structure). These projections to the same-area or another-area patches are created in a random manner — the probability that the projection will be created is subject to the Gaussian density function centered in the patch center. In exactly the same way other e-cells can send projections to the e-cell of interest. Secondly, the interaction of this e-cell with i-cells in the same area will be considered (note that i-cells interact only with the e-cells from the same area). Each i-cell corresponds exactly one e-cell and sends its projection only to this e-cell. This i-cell gets projections from each cell in 5x5 patch of e-cells around the e-cell of interest. The i-cell sums up inputs from the patch of the e-cells and proportionally to this inhibits the e-cell of interest, which reflects the local inhibition mechanism in this model. Accordingly, each e-cell sends projections to the i-cells to inhibit its neighboring e-cells. In the baseline model, I will use this micro-connectivity pattern without changes.

3.1.3 Dynamics of the model. Firstly, dynamics of the e-cells changes will be considered (see **Fig. 3**). Changes in membrane potential of the e-cell depend on the current membrane potential, summed net input to this e-cell (this component will be explained further) and white noise[3] specific to the cell (see **(E1)**). White noise is an important biological feature of this model. It is presented only in the e-cell and reflects the averaged spontaneous activity of the excitatory neurons. E-cell output depends on the value of its membrane potential (depicted by the arrow **A1**) and the threshold (see **(E2)**), where the threshold is sensitive to the averaged recent activity (output) of the e-cell — the higher recent activity is the higher the threshold for the current output will be (see **(E2.1)** and **(E2.2)**). An important mechanism in this model is the global inhibition of the e-cells. It is based on the sum of the e-cells' output from the same area as the e-cell under consideration (see **(E3)**) — the higher activity of all other e-cells in the area is the stronger inhibition of the e-cell of interest will be. This

[3] See the suggested parameters in the Appendix

global inhibition influences the net input to the e-cell (depicted by the arrow **A3**) which in turn shapes the membrane potential. The net input to the e-cell consists of the sum of all excitatory and inhibitory postsynaptic potentials[4] (E/IPSPs) and scaled global inhibition component. EPSP from the e-cell *x* to the e-cell *y* (as well as IPSP from the i-cell *x* to the e-cell *y*) is defined as the output of the cell *x* multiplied by the weight of the synaptic connection from the cell *x* to the cell *y* (weights initialization and the rule according to which weights change will be explained further).

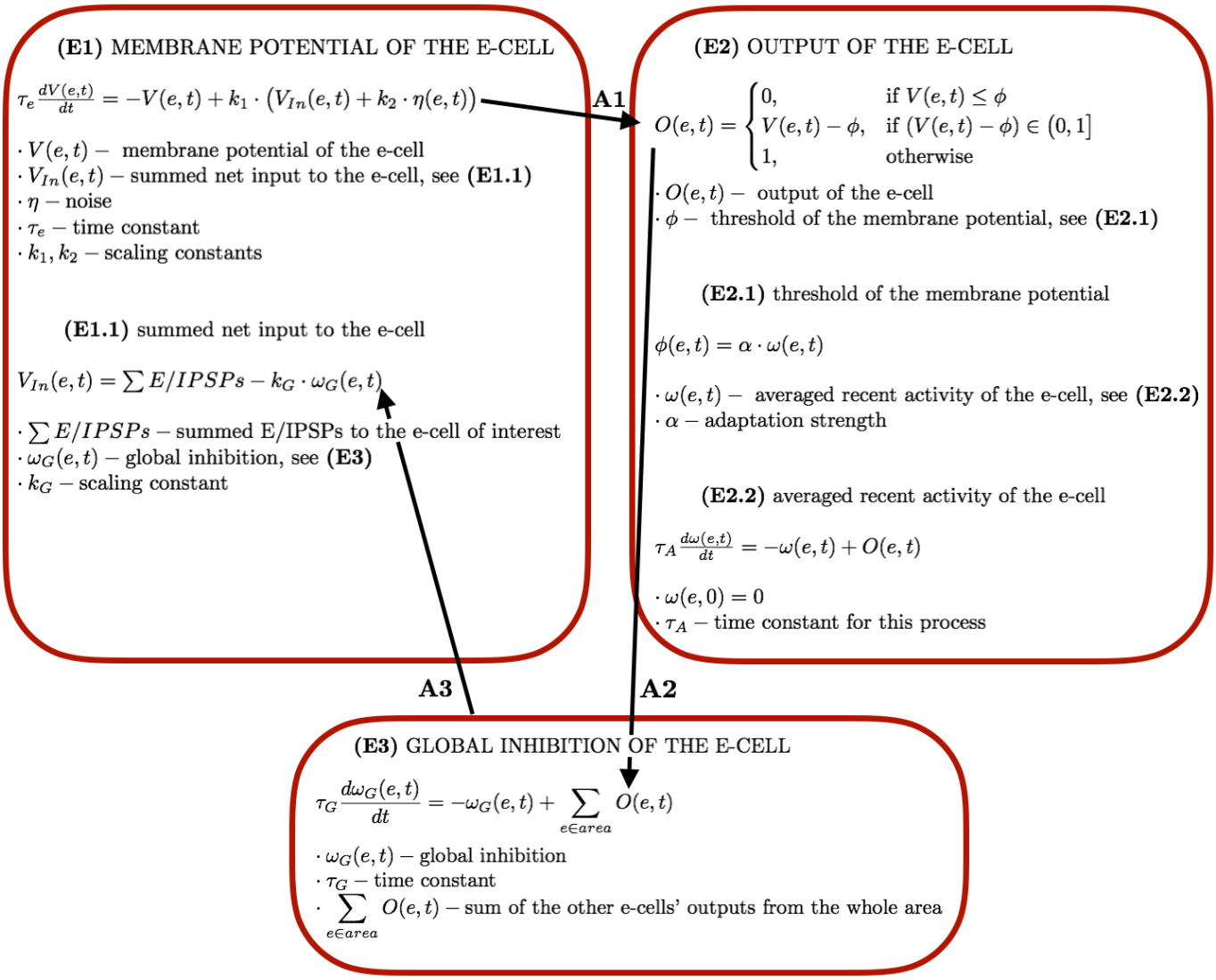

Figure 3. Dynamics of the e-cell.

[4] Each e-cell gets only one IPSP from the corresponding i-cell, which is included into this sum with a negative sign

Secondly, the dynamics of the i-cells changes will be considered (see **Fig. 4**). In this case, changes in the membrane potential of the cell depend on the current membrane potential and summed net input to this i-cell (see **(I1)**). The net input to the i-cell consists only of EPSPs from the 5x5 e-cells patch (see **(I1.1)**). I-cell output depends only on the value of its membrane potential (depicted by the arrow **A4**) — if the membrane potential is greater than zero, inhibition will occur (see **(I2)**).

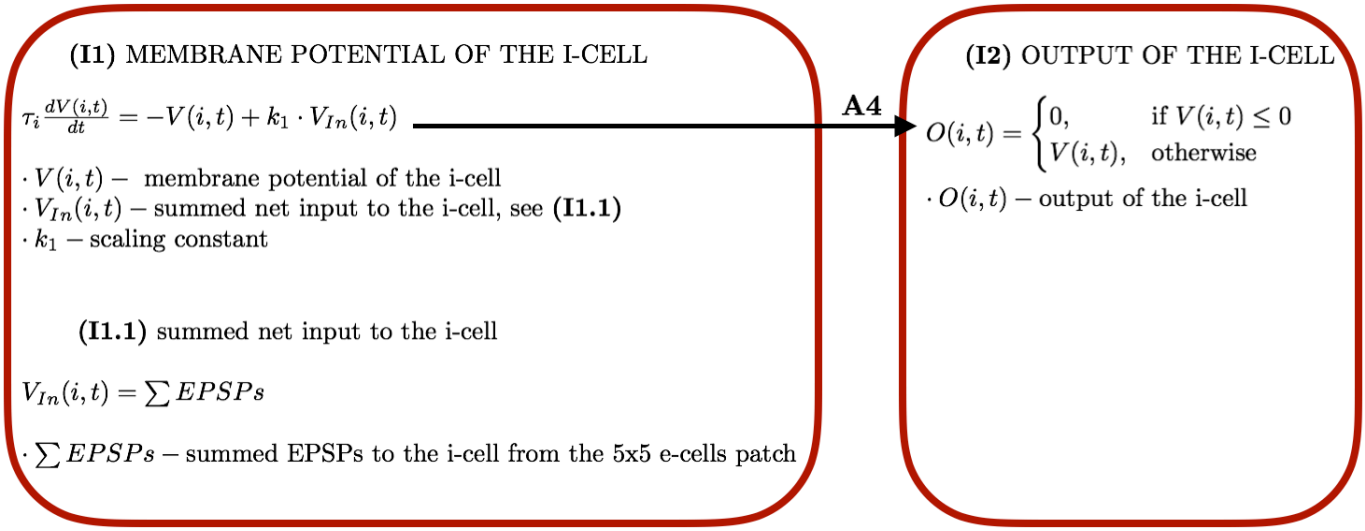

Figure 4. Dynamics of the i-cell.

Finally, the initialization of the synaptic weights and the rule according to which they change during the learning phase will be discussed. Initially, random weights[5] are assigned to all established connections (recall that connections' creation is also random and is subject to the Gaussian density function). Further, Hebbian learning takes place (see **Fig. 5**).

If high enough pre-synaptic activity (cell's output) corresponds to high enough post-synaptic activity (membrane potential), the synaptic weight strengthening (long-term potentiation) occurs, subject to the rule "fire together — wire together". If low pre-synaptic activity corresponds to high post-synaptic activity, the synaptic weight

[5] See the suggested parameters in the Appendix

weakening (long-term depression) occurs, subject to the rule "out of sync — out of link" (Artola & Singer, 1993).

(L1) HEBBIAN LEARNING MECHANISM

$$w_{t+1}(x,y) = \begin{cases} w_t(x,y) + \Delta w, & \text{if } O(x,t) > \theta_{pre} \text{ and } V(y,t) > \theta_{post} \\ w_t(x,y) - \Delta w, & \text{if } O(x,t) \leq \theta_{pre} \text{ and } V(y,t) > \theta_{post} \\ w_t(x,y), & \text{otherwise} \end{cases}$$

- $w_t(x,y)$ – weight of the synapse from the cell x to the cell y at time t
- Δw – change of the synaptic weight
- θ_{pre} – threshold for the presynaptic activity
- θ_{post} – threshold for the postsynaptic membrane potential

Figure 5. Hebbian learning mechanism

These dynamic patterns will be used in the baseline model without changes, however, exact model parameters will be specified during the next stage of the research.

3.1.4 Semantic categories. In the original model, two semantic categories were suggested — action-related words and object-related words (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018). The model was taught to acquire object-related semantics by co-experiencing and associating auditory, articulatory and visual patterns — inputs to A1, $M1_i$ and V1 were presented simultaneously during the learning phase (see **Fig. 6A**). To teach action-related semantics, inputs were provided to auditory (A1), articulatory ($M1_i$) and motor ($M1_L$) primary areas (see **Fig. 6B**). This mechanism of co-experiencing of the different perceptions to learn semantics reflects the idea about child teaching — for example, mom teaches her child to name "cat", pronouncing the word "cat", asking her child to repeat and at the same time pointing to the cat.

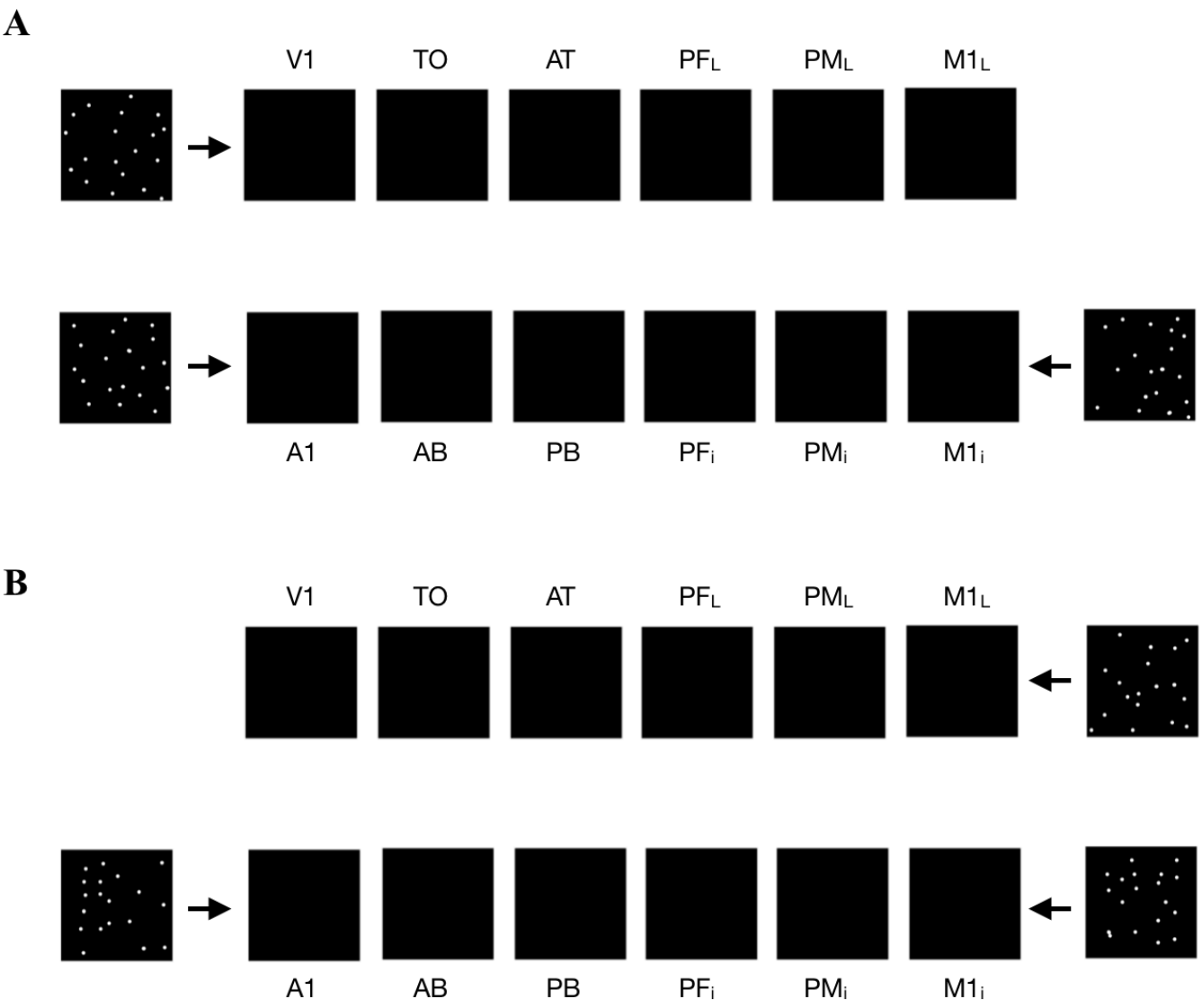

Figure 6. **A.** Example of the acquisition of object-related semantics via inputs to auditory (A1), articulatory ($M1_i$) and visual (V1) primary areas. **B.** Example of the acquisition of action-related semantics via inputs to auditory (A1), articulatory ($M1_i$) and motor ($M1_L$) primary areas.

In the SD model, I want to show the progressive loss of semantic knowledge which starts from deterioration of specific semantic details (sub-categories) and spreads to more general ones (categories). Therefore, two sub-categories in each of the main semantic category (object-related and action-related words) will be modeled — for instance, “birds” and “fishes” can be two sub-categories of the object-related words.

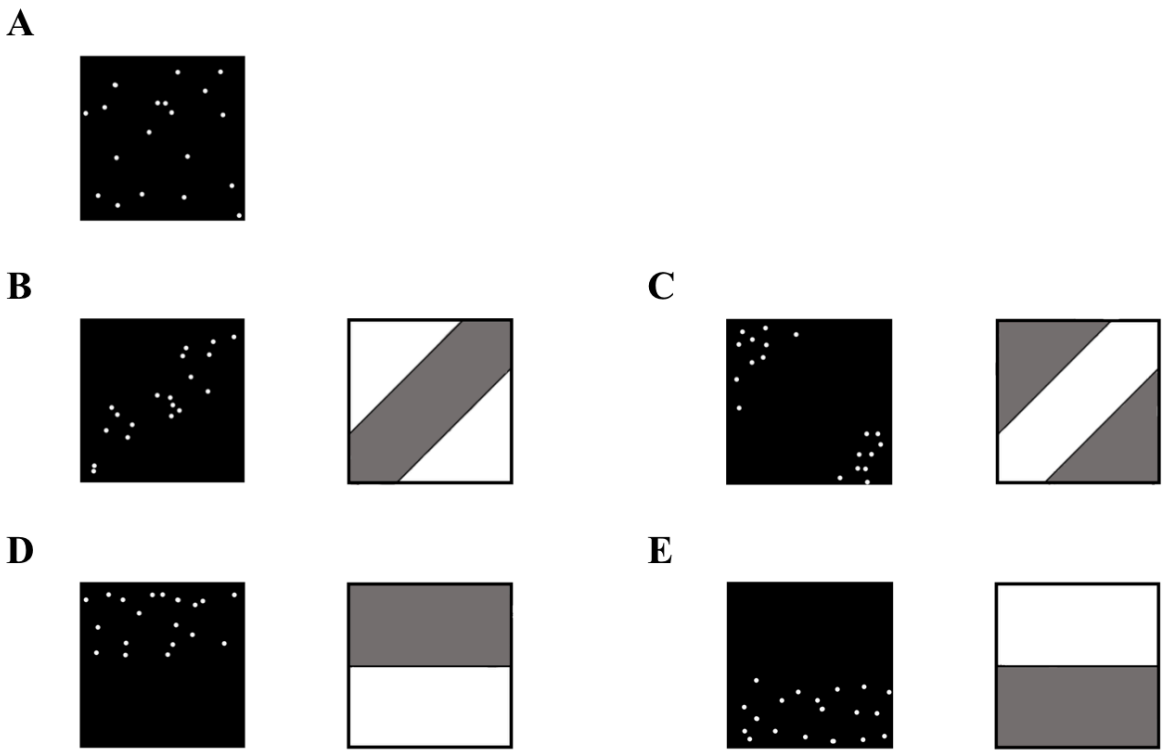

Figure 7. Examples of random inputs during concept acquisition. **A.** Example of the input to the A1 or $M1_i$. **B.** Sub-category #1 of the object-related words. Left: example of the input to the V1; right: for this sub-category, random 19-cell pattern will be distributed only in the depicted shaded area around the diagonal, which is the half of the whole area. **C.** Sub-category #2 of the object-related words. Left: example of the input to the V1; right: for this sub-category, random 19-cell pattern will be distributed only in the depicted shaded area in two corners, which is the half of the whole area. **D.** Sub-category #1 of the action-related words. Left: example of the input to the $M1_L$; right: for this sub-category, random 19-cell pattern will be distributed only in the depicted upper half shaded area. **E.** Sub-category #2 of the action-related words. Left: example of the input to the $M1_L$; right: for this sub-category, random 19-cell pattern will be distributed only in the depicted lower half shaded area.

Acquisition of the sub-category #1 of object-related words will start from random 19-cell inputs[6] simultaneously to A1, $M1_i$ and V1. Whereas A1, $M1_i$ input

[6] This idea of the cells activity pattern was used in the original model in the same manner (Garagnani & Pulvermüller, 2016)

will be randomly distributed through all area cells (see **Fig. 7A**), V1 input will be randomly distributed around the diagonal of the area (see **Fig. 7B**). For the sub-category #2 of object-related words, A1, $M1_i$ input will be randomly distributed through all area cells (see **Fig. 7A**), while V1 input will be randomly distributed in the two corners of the area (see **Fig. 7C**). This distributed activity in the primary visual cortex reflects the idea that objects from the one sub-category have similar shapes, therefore, activate primary visual cortex in a similar way; and, on the contrary, objects from the different sub-categories have different shapes, therefore, activate primary visual cortex differently. Analogous idea to distinguish between objects shapes was used to teach the modification of the original model to associate objects and actions connected to these objects (Adams, Wennekers, Cangelosi, Garagnani, & Pulvermüller, 2014). Acquisition of the sub-category #1 of action-related words will start from the random 19-cell inputs simultaneously to A1, $M1_i$ and $M1_L$. A1, $M1_i$ input will be randomly distributed through all area cells (see **Fig. 7A**), $M1_L$ input will be randomly distributed in the upper half of the area (see **Fig. 7D**). For sub-category #2 of action-related words, $M1_L$ input will be randomly distributed in the lower half of the area (see **Fig. 7E**). This distributed activity resonates with the topography of the primary motor cortex — where hand movements are coded by the lower part of the area, and leg movements are coded by its upper part (Hauk & Pulvermüller, 2004).

3.1.5 Learning procedure. Five word patterns for each of the four sub-categories will be generated meaning that the model will learn 20 different word patterns. Each word pattern will consist of inputs to three out of four primary areas, as discussed above. During the learning phase, the model will be taught each word pattern by the simultaneous presenting of the inputs to three out of four different modalities. Each word pattern will be presented during 3000 trials. One trial will last 16 time steps. Next trial will start as soon as activity in the network fall below the threshold, trials will be randomly shuffled (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018). It is important to note that white noise will be presented to all four primary areas, as it will reflects the averaged spontaneous activity of the areas.

In addition, a random 19-cell noise pattern will be provided to the not-involved primary area — V1 for action-related words and $M1_L$ for object-related ones. This noise pattern will be changed in each trial, so it differs from the stable 19-cell pattern to the primary areas which are involved in the semantic acquisition (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018).

3.1.6 Expected results. Twenty model exemplars (with the structure described above) will be initialized and will undergo the learning procedure (which is described above): this will reflect the sample of the modeled participants.

It was shown in the original work, that, after the learning phase, a particular subset of e-cells (cell assembly — CA) distributed in the deferent areas will build stronger connections and produce strong robust activity in response to the input word patterns (Garagnani & Pulvermüller, 2016). These e-cells provide a basis for semantic knowledge representations. CAs differ for different word patterns.

To evaluate which cells will be included into the CA for the particular word *W* after the learning phase, the following procedure will be performed: (1) one time-step stimulation of the auditory and articulatory primary areas with the word *W* input patterns; (2) calculating the time-averaged output of each e-cell in the model for 15 time-steps; (3) revealing the highest averaged e-cell output in each area; (4) each e-cell, which shows at least half of the highest output in its area, is included into the CA for word *W*. This procedure is suggested by the authors of the original work (Garagnani & Pulvermüller, 2016). Number of cells which are included into the CAs, will be averaged for each area across model exemplars ("modeled participants") and across all the words from the same sub-category.

The following results regarding CAs' distribution are expected. First of all, for object-related words higher activity (higher averaged number of cells in CAs) is expected in the primary and secondary visual areas (V1 and TO, accordingly) than in the primary and secondary motor areas ($M1_L$ and PM_L, accordingly). The opposite pattern is expected for the action-related words. Second, higher activity is expected in the "central areas" of the visual and motor zones than in the primary and secondary ones; and no difference in the "central areas" activity after object-related words

presentation and activity after action-related words presentation is expected. The results described above were shown in the original model (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018). They demonstrate that the model learns to discriminate between different categories of semantics on basis of sensory-motor representations and at the same time multimodal "connector hubs" appear, which is consistent with the empirical results.

Third, for the sub-category #1 of object-related words, higher activity in the area around the diagonal in V1 than in the corners is expected (and the opposite pattern for the sub-category #2 of object-related words). Correspondingly, for the sub-category #1 of action-related words, higher activity in the upper half of the $M1_L$ area is expected (and the opposite pattern for the sub-category #2 of action-related words). These results will reflect that the model learns how to distinguish between different sub-categories inside one category.

3.2 Description of the Semantic Dementia Model

In this part of the paper, I will suggest changes to the baseline model which reflect neuroanatomical deterioration during semantic dementia. Initially, the baseline model will be built (as described above) and undergo the learning procedure (as described above). Then, gradual changes in the micro-structure will occur. Note that changes will occur only in the micro-structure of the baseline model (macro-structure and predefined dynamics will not be changed) and only after the learning phase. This reflects the idea that SD develops after the semantic system is already established, as SD typically develops in elderly people.

3.2.1 Changes in the micro-structure. As discussed in the part *Semantic dementia* in the *Literature review*, SD is characterized by the progressive white and gray matter degradation in the ATL. In this model, progressive deterioration of white matter (WM) will be reflected by the gradual loss of connections to e-cells in the AT area from other areas' e-cells and gradual loss of connections from e-cells in the AT area to other areas' e-cells. Three levels of deterioration will be implemented (these are current estimates — they will be adjusted during the model testing phase): mild — 25% of randomly chosen connections to and from AT area will be lost; medium —

50% of randomly chosen connections will be lost, and heavy — 75% of randomly chosen connections will be lost. Progressive deterioration of grey matter (GM) will be reflected by the gradual loss of e-cells in AT area. Three levels will be implemented (these are current estimates — they will be adjusted during the model testing phase): mild — 25% of randomly chosen e-cells in AT area will be lost; medium — 50%, and heavy — 75%. Different combinations of simultaneous deterioration of WM and GM will be implemented as well.

3.2.2 Expected results. As in the baseline model, twenty model exemplars will be initialized (with preserved micro-connectivity) and undergo the learning procedure. After the learning phase, AT deterioration (either the loss of connections or loss of e-cells or both) will be implemented, as discussed above. Then, CAs for each word will be evaluated using the algorithm discussed in the baseline model description.

The following results regarding CAs' distribution are expected. With mild deterioration (either of connections or of e-cells or both), discrimination between sub-category #1 and sub-category #2 for both semantics will be disturbed. For example, similar activity is expected around the diagonal and in the corners of the V1 area when either sub-category #1 or sub-category #2 words are presented. With more severe deterioration, discrimination between action-related words and object-related words will be disturbed — similar activity is expected in primary and secondary visual areas (V1 and TO) and in primary and secondary motor areas ($M1_L$ and PM_L) when either action- or object-related words are presented. These results reflect the empirical evidence of the progressive loss of semantic knowledge which starts from deterioration of specific semantic details and spreads to more general ones.

I have no predefined hypothesis regarding what will lead to more severe performance — connections or cells deterioration. I also have no predefined hypothesis regarding the results of the interaction between two loss types. This part of research will be exploratory, as the data and the theories we have at the moment do not allow us to make any concrete predictions.

Acknowledgements

I would like to thank Dr. Max Garaniani, who kindly shared the source code of the model with me and helped me with the work.

I also thank Alexey Guzey for inspiration and editing.

References

Adams, S. V., Wennekers, T., Cangelosi, A., Garagnani, M., & Pulvermüller, F. (2014, December). Learning visual-motor cell assemblies for the iCub robot using a neuroanatomically grounded neural network. In *2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)* (pp. 1-8). IEEE.

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Aguilar, M., … & Raizada, R. D. (2016). Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, *27*(9), 4379-4395.

Artola, A., & Singer, W. (1993). Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. *Trends in neurosciences*, *16*(11), 480-487.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577-660.

Brambati, S. M., Amici, S., Racine, C. A., Neuhaus, J., Miller, Z., Ogar, J., ... & Gorno-Tempini, M. L. (2015). Longitudinal gray matter contraction in three variants of primary progressive aphasia: A tenser-based morphometry study. *NeuroImage: Clinical*, *8*, 345-355.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, *33*(3-4), 130-174.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in cognitive sciences*, *15*(11), 527-536.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767-2796.

Caramazza, A., Anzellotti, S., Strnad, L., & Lingnau, A. (2014). Embodied cognition and mirror neurons: a critical assessment. *Annual review of neuroscience*, *37*, 1-15.

Chen, L., Ralph, M. A. L., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature human behaviour*, *1*(3), 0039.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., & Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature*, *380*(6574), 499.

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, *22*(3-4), 455-479.

Garagnani, M., & Pulvermüller, F. (2016). Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs. *European Journal of Neuroscience*, *43*(6), 721-737.

Garagnani, M., Wennekers, T., & Pulvermüller, F. (2007). A neuronal model of the language cortex. *Neurocomputing*, *70*(10-12), 1914-1919.

Garagnani, M., Wennekers, T., & Pulvermüller, F. (2008). A neuroanatomically grounded Hebbian−learning model of attention–language interactions in the human brain. *European Journal of Neuroscience*, *27*(2), 492-513.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335-346.

Hauk, O., & Pulvermüller, F. (2004). Neurophysiological distinction of action words in the fronto−central cortex. *Human brain mapping*, *21*(3), 191-201.

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory*, *3*(3-4), 463-495.

Hodges, J. R., Mitchell, J., Dawson, K., Spillantini, M. G., Xuereb, J. H., McMonagle, P., … & Patterson, K. (2009). Semantic dementia: demography, familial factors and survival in a consecutive series of 100 cases. *Brain*, *133*(1), 300-306.

Hodges, J. R., & Patterson, K. (2007). Semantic dementia: a unique clinicopathological syndrome. *The Lancet Neurology*, *6*(11), 1004-1014.

Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia: Progressive fluent aphasia with temporal lobe atrophy. *Brain*, *115*(6), 1783-1806.

Hoffman, P., & Lambon Ralph, M. A. (2011). Reverse concreteness effects are not a typical feature of semantic dementia: evidence for the hub-and-spoke model of conceptual representation. *Cerebral Cortex*, *21*(9), 2103-2112.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453.

Joubert, S., Vallet, G. T., Montembeault, M., Boukadi, M., Wilson, M. A., Rouleau, I., & Brambati, S. M. (2017). Comprehension of concrete and abstract words in semantic variant primary progressive aphasia and Alzheimer's disease: a behavioral and neuroimaging study. *Brain and language*, *170*, 93-102.

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805-825.

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of physiology-Paris*, *102*(1-3), 59-70.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, Vol. 1

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*(7), 788-804.

Montembeault, M., Brambati, S. M., Gorno-Tempini, M. L., & Migliaccio, R. (2018). Clinical, anatomical, and pathological features in the three variants of primary progressive aphasia: a review. *Frontiers in neurology*, *9*.

Murre, J. M., Graham, K. S., & Hodges, J. R. (2001). Semantic dementia: relevance to connectionist models of long-term memory. *Brain*, *124*(4), 647-675.

Patterson, K., & Erzinçlioğlu, S. W. (2008). Drawing as a 'window' on deteriorating conceptual knowledge in neurodegenerative disease.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976.

Pobric, G., Jefferies, E., & Ralph, M. A. L. (2007). Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rTMS in normal participants. *Proceedings of the National Academy of Sciences*, *104*(50), 20137-20141.

Pulvermüller, F. (2001). Brain reflections of words and their meaning. *Trends in cognitive sciences*, *5*(12), 517-524.

Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, *21*(3), 793-797.

Quillan, M. R. (1966). *Semantic memory* (No. SCIENTIFIC-2). BOLT BERANEK AND NEWMAN INC CAMBRIDGE MA.

Quillian, M. R. (1969). The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM*, *12*(8), 459-476.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42

Ralph, M. L., Lowe, C., & Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain*, *130*(4), 1127-1137.

Rogers, T. T., Patterson, K., Jefferies, E., & Ralph, M. A. L. (2015). Disorders of representation and control in semantic cognition: Effects of familiarity, typicality, and specificity. *Neuropsychologia*, *76*, 220-239.

Rogers, T. T., Ralph, L., Matthew, A., Garrard, P., Bozeat, S., McClelland, J. L., … & Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review*, *111*(1), 205.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological review*, *81*(3), 214.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences*, *11*(1), 1-23.

Spinelli, E. G., Mandelli, M. L., Miller, Z. A., Santos‐Santos, M. A., Wilson, S. M., Agosta, F., ... & Henry, M. L. (2017). Typical and atypical pathology in primary progressive aphasia variants. *Annals of neurology*, *81*(3), 430-443.

Tomasello, R., Garagnani, M., Wennekers, T., & Pulvermuller, F. (2018). A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity. *Frontiers in computational neuroscience*, *12*, 88.

Tomasello, R., Wennekers, T., Garagnani, M., & Pulvermüller, F. (2019). Visual cortex recruitment during language processing in blind individuals is explained by Hebbian learning. *Scientific reports*, *9*(1), 3579.

Ueno, T., Saito, S., Rogers, T. T., & Ralph, M. A. L. (2011). Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, *72*(2), 385-396.

Yi, H. A., Moore, P., & Grossman, M. (2007). Reversal of the concreteness effect for verbs in patients with semantic dementia. *Neuropsychology*, *21*(1), 9.

Appendix

Table 1

Suggested parameters in the model[7]

$\tau_e = 2.5$	time constant for membrane potential of the e-cell (E1)
$\tau_i = 5$	time constant for membrane potential of the i-cell (I1)
$\tau_G = 12$	time constant for global inhibition (E3)
$\tau_A = 10$	time constant for averaged recent activity of the e-cell (E2.2)
$k_1 = 0.0008$	scaling constant (E1)
$k_2 = 25 \cdot \sqrt{48}$	scaling constant (E1.1)
$k_G = 65$	scaling constant (E1.1)
$\alpha = 0.01$	adaptation strength (E2.1)
$\eta \sim U[-0.5, 0.5]$	noise distribution (E1)
$w_{initial} \sim U[0, 0.1]$	distribution of initial weights
$\Delta w = 0.0008$	change of the synaptic weight during Hebbian learning (L1)
$\theta_{pre} = 0.05$	threshold for presynaptic activity (L1)
$\theta_{post} = 0.15$	threshold for postsynaptic activity (L1)

[7] These parameters are approximate for now and can be changed during the model testing