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"

Final qualifying work – MASTER THESIS

«Biologically-Inspired Neurocomputational Model of Semantic Dementia»

Student
Kuptsova, Anastasia
Last name, First name Middle name

_____
Signature

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

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

Lecturer, PhD
Position, Academic degree
Garagnani, M.
Last, F. M./O.

Moscow, 2020

Table of Contents

Chapter 1. Introduction

Semantic knowledge is the knowledge about meaning of concepts and words (Quillan, 1966; Kiefer & Pulvermüller, 2012). We constantly use it to behave properly in different situations, to express our thoughts, to understand others, and so on. People with semantic memory impairments experience enormous difficulties in everyday life — they can fail to recognize or produce speech (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 can fail to understand and explain the source of the pain that they feel (Montembeault et al., 2018). This combined evidence means that semantic knowledge constitutes the core units of our thoughts and behavior. Revealing the semantic system organization would perhaps allow us to take a step toward solving the eternal mystery of human cognition.

Early theories assumed that the semantic system consists of symbolic representations of real-world entities in the brain — for instance, that each semantic concept is just a node in the network (Quillan, 1969; Collins & Loftus, 1975) or that each concept could be represented by the features list (Smith, Shoben, & Rips, 1974) — and to use our semantic memory we need to manipulate these symbols. The problem with these approaches was in the following: how can our real-world experience, which we need to acquire concepts, be transduced into symbols that are totally distinct from this sensory-motor experience (Harnad, 1990); in other words, how can our brain acquire and store absolutely amodal representations of concepts (Barsalou, 1999). In 1990s the first evidence of concepts being grounded in sensory and motor modalities appeared, soon followed by theories utilizing it (Barsalou, 1999; Pulvermüller, 2001; Damasio, Grabowski, Tranel, Hichwa, & Damasio, 1996). However, different research groups have continued to debate the degree of grounding. For instance, in 2016 the special issue of the *Psychonomic Bulletin & Review* journal was devoted to the concept grounding problem (for review see Barsalou, 2016). Some groups posit that the sensory-motor system plays only a secondary role in semantic representations (Mahon & Caramazza, 2008; Mahon, 2015) and that it is

important to have one higher-order brain area where all concepts are stored in the amodal way (Rogers et al., 2004; Patterson, Nestor, & Rogers, 2007). Other groups posit that when we process a concept, our brain almost simulates the perceptual experience of the interaction with this concept (Gallese & Lakoff, 2005). Finally, there are groups that propose hybrid theories where both the necessary grounding in the perceptual experience and higher-order convergence zones to mediate the experience from different modalities are suggested (Kiefer & Pulvermüller, 2012; Garagnani & Pulvermüller, 2016; Binder & Desai, 2011; Binder, 2016; Barsalou, 2016).

While current evidence speaks in favor of hybrid theories (Kiefer & Pulvermüller, 2012; Meteyard, Cuadrado, Bahrami, & Vigliocco, 2012; Binder, Desai, Graves, & Conant, 2009; Barsalou, 2016), there still exist arguments against these theories, challenging the role of embodiment. Three main arguments are mostly raised by the authors of the secondary embodiment theories (Mahon & Hickok, 2016; Leshinskaya & Caramazza, 2016; Patterson et al., 2007; Machery, 2016). The first argument states that anterior temporal lobe (ATL) degradation in semantic dementia (SD) patients proves the amodal core of concept representations (Patterson et al., 2007; Machery, 2016). The second argument states that, if there exist patients with lesions in the sensory-motor system who have no problems with semantic knowledge, then concepts' grounding in the sensory-motor system is not important (Mahon & Hickok, 2016; for review see Barsalou, 2016). The third argument states that, if congenitally blind individuals do not differ from sighted individuals in the visually-related concepts they formed, then we can acquire exactly the same conceptual information even without first-hand visual experience (Mahon & Hickok, 2016; Leshinskaya & Caramazza, 2016). Authors of hybrid theories show that arguments above do not contradict the hybrid idea (Kiefer & Pulvermüller, 2012; Barsalou, 2016), however, these arguments continue to be used against it.

In this paper, we work on the semantic dementia (SD) model. SD is the core disease affecting semantic memory (Hodges & Patterson, 2007; Montembeault et al., 2018; Spinelli et al., 2017). We augment an existing biologically-inspired neuronal

model of the semantic system (Garagnani & Pulvermüller, 2016; Tomasello, Garagnani, Wennekers, & Pulvermuller, 2018) that was built in the hybrid paradigm with the SD mechanism and compare the model's prediction with the SD patients' data. Our model also generates additional predictions that can be further tested in the experiments with SD patients.

We want to highlight the following goals of this paper:

(1) SD is the most important disease of the semantic system — it can be crucial to have a biologically accurate model of SD to understand its mechanism and overall semantic system organization better. Thus, the main goal of this study is to build such a model.

(2) Although the baseline model (to which we add our SD mechanism) is one the most biologically precise models of the semantic system that exist to date (see argumentation in section *2.2 Approaches to Semantic System Modeling*), still, a tremendous amount of things could be improved. For instance, the bridge between the real data and bio-inspired architecture of the model is currently missing. However, before shifting it to the real data domain, important qualitative results should be checked for this model, in order to understand whether this "best" model is worthwhile to work on or if there are fatal problems in it. Therefore, another goal of the current research is to simulate the SD mechanism in the baseline model, in order to check the qualitative predictions of it and to understand whether we should improve this model further or not.

This paper is organized as follows. In the second chapter, we discuss theories of semantic system organization, neurocomputational attempts to model the semantic system and data from semantic dementia patients. Then, in the third chapter, we describe the baseline model (where we add our SD mechanism), details of the SD mechanism implementation, and our predictions regarding this mechanism. In the fourth chapter, we describe our results of the replication of the baseline model and SD implementation there. Finally, in *Discussion*, we examine our results and check the fulfillment of the initial goals which we set. Additionally, we broadly discuss how

data that raises main arguments against hybrid theories can be alternatively explained by our SD model and by the baseline model.

Chapter 2. Literature Review

2.1 Theories of Semantic System Organization

2.1.1 Early theories. Starting from the second part of the 20th-century, scientists have been trying to reveal the mechanisms and principles of human cognition. Back then, scientists also actively tried to use these “cognitive” mechanisms in the construction of artificial intelligence. They believed they could build AI very soon and make it very biologically-grounded. For this reason, early cognitive theories were attempted to simplify the explanations of the cognitive mechanisms to the point where then-existing computers could simulate them (Barsalou, 1999; Meteyard et al., 2012).

The symbolic approach to semantic representation was one of such early cognitive theories. It states that the perceptual sensory-motor experience which we gain while interacting with the real world entities or events is transduced into symbolic representations and then these representations are used in different aspects of cognition. Quillan (1969) built a computational program for the mechanical extraction of meaning from text. He argued that the theory that underlies this program could be related to semantic processing by humans as well. Quillan (1969) and Collins and Loftus (1975) presented the semantic memory model as a network with distinct nodes, which symbolize concepts, and links between the related nodes. Different types of links in this network reflect the different relationships between concepts, for instance, the 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 et al., 1974). Authors 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 trees, green, and so on. Here, *green* is a less common feature than *round*.

Another branch of 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, the semantic system could be defined as a network with neuron-like units, some of which are connected via links. These links have weights that are dynamically changing under the input patterns and the learning rule for the network. In this approach, concepts could be represented as a 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, Lambon Ralph, & Rogers, 2017).

2.1.2 Symbol grounding problem. In both the symbolic (Quillan, 1969; Collins & Loftus, 1975; Smith, 1974) and the early connectionist (McClelland, et al., 1986; Smolensky, 1988) approaches, semantic representations were amodal — distinct from the sensory-motor experience and from the sensory-motor 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 amodal representations from the real-world entities and shouldn't meaningfulness imply at least some grounding in the interaction with the real world? This question is known as the "symbol grounding problem" (Harnad, 1990). In 1990s, first pivotal works about semantic grounding in the sensory-motor system appeared (Barsalou, 1999; Pulvermüller, 2001; Damasio et al., 1996). It was shown in patients with focal cortical lesions and in healthy people with the help of TMS that comprehension and retrieval of different categories of concepts (e.g. animals versus tools) correlate with different cortical sites (e.g. occipital and inferior temporal areas versus premotor areas, accordingly) (for review see Barsalou, 1999; Pulvermüller, 2001). It was assumed that the neuronal network in the sensory and motor areas, which was built from the sensory-motor associations during concept acquisition, could be the basis for semantic representations. To illustrate this idea, the concept "apple" can be

represented by distributed sensory-motor associations like the visual, gustatory, and olfactory perception of an apple, the auditory perception of its name “apple”, the motor knowledge of how to grasp an apple, and so on. Barsalou (1999) proposed that, during 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 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?

2.1.3 Modern amodal (secondary embodied) theories. One extreme answer to the question of embodiment’s importance proposes that sensory-motor activation plays only the secondary or complementary role for the semantic system while the amodal core is much more important. 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 dynamics of activation between and within perceptual and conceptual systems should be studied meticulously. Hypothetically, the visual perception of an apple can (1) directly activate fully distributed sensory-motor representation of the apple concept; or (2) first, activate distributed sensory-motor representation, which, in turn, activates the amodal representation of the concept of apple; or (3) first, activate amodal representation, which, in turn, activates distributed sensory-motor representation of the concept of apple. This third dynamics is assumed by Mahon and Caramazza (2008) in their "grounding by interaction" hypothesis or by Mahon (2015) in the “default explanation”. They argue that only 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, & 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 names, shapes, colors, tastes, and so on, however, we somehow generalize them into the one fruit category. To integrate all attributes of the concept, the higher-order amodal zone in the anterior temporal lobe (ATL) — the semantic 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).

The main problem of the secondary embodied theories (this name was borrowed from Meteyard et al., 2012) is the lack of reasonable theoretical descriptions of the acquisition mechanism that transfer concepts from reality to their amodal neuronal representations (Barsalou, 2016). If we agree that one of the ways to learn new words is to co-experience the sensory and language inputs where primary sensory areas play a very important role — child is hearing the word "dog" and seeing her mom pointing to a dog — then what is the mechanism of detachment of representations from this sensory experience and of mapping them into the amodal format and why does this mechanism need to detach these representations at all?

Moreover, these amodal representations cannot fully explain the data about the correlation between the activation of the different cortical sites and the comprehension of the specific categories of concepts. Patients with lesions in the motor and premotor cortex often 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 (TMS) of the hand motor areas could enhance comprehension of the hand-related words, while

stimulation of the leg areas could do a similar thing with the leg-related words (Pulvermüller, Hauk, Nikulin, & Ilmoniemi, 2005).

To show the contradiction between these secondary embodiment theories and empirical data, let us assume the following situation: the patient has a lesion in the 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 propagates 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, in this amodal 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.

2.1.4 Strong embodiment theories. Another extreme answer to the question about the extent of the 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 imagining an action and performing an action share the same neuronal sites. Then they posit that imagination of an action and knowledge about this action share the same neuronal sites, as it is hard to imagine that we cannot imagine some action but can have knowledge about it. Thus, they conclude that taking an action and having the knowledge about this action share the same neuronal sites. Furthermore, similar logic (and the appropriate empirical findings) can be used to prove that objects representations are grounded in the sensory-motor system. Therefore, the authors infer that the structure of the sensory-motor system is sufficient for semantic representations and that the representation of a concept is the simulation of the perceptual experience of this concept.

The central flaw of the strong embodiment theories 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 (and therefore from the

perceptual experience). The review of the 120 fMRI studies shows that outside of the sensory-motor system there exist multiple zones, for instance, in the lateral and ventral temporal areas and in the inferior parietal area, which are important for general semantic processing (Binder et al., 2009). Moreover, patients with semantic dementia have progressive anterior temporal lobe (ATL) atrophy, which correlates with the semantic knowledge deterioration, even though ATL is not located in the sensory-motor system (Hodges & Patterson, 2007; Montembeault et al., 2018; Spinelli et al., 2017).

The evidence above suggests that we need something outside of the perceptual system to store conceptual knowledge, therefore, the idea that conceptual knowledge is just a perceptual simulation in the sensory-motor system is unlikely.

2.1.5 Hybrid theories. In between these extreme answers to the question of embodiment, there exist hybrid theories that assume both the essential grounding of concepts in the sensory-motor perceptual system and the need for the higher-order convergence zones. Pulvermüller and colleagues (Kiefer & Pulvermüller, 2012; Garagnani & Pulvermüller, 2016) suggest that concept representations are distributed through modality-specific sensory-motor areas, which play an important role in concept acquisition and storage, and "connector hubs", which help with mediation and connection between modalities but do not work as the core amodal storage facilities for concepts, as was proposed by the authors of the secondary embodiment theories (Mahon & Caramazza, 2008; Mahon, 2015; Rogers et al., 2004; Patterson et al., 2007). Pulvermüller and colleagues argue that this hypothesis follows directly from the perceptual origin of the semantic knowledge acquisition and the neurobiological properties of the brain (Pulvermüller, 2001; Garagnani & Pulvermüller, 2016). When we acquire a concept, it is experienced by different perceptual modalities simultaneously (e.g. mom teaches her child "This is an apple" showing 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). Furthermore, 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 proposed by Binder and Desai (2011) and in the “CCR theory” proposed by Binder (2016), authors suggest 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 the context, the concept familiarity and the 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.

There are three main arguments against the importance of embodiment in hybrid theories (as hybrid theories claim the essential grounding of concepts in the sensory-motor perceptual system) that appear mostly from the authors of the secondary embodiment theories (Mahon & Hickok, 2016; Leshinskaya & Caramazza, 2016; Patterson et al., 2007; Machery, 2016). However, these three arguments do not disprove hybrid theories, quite the opposite, they give hybrid theories an opportunity to be challenged and to survive. The arguments have been widely discussed and refuted by the authors of hybrid theories (Kiefer & Pulvermüller, 2012; Barsalou, 2016).

The first argument states that SD patients’ problems reveal the amodal core of the concept representations, which diminishes the importance of concepts embodiment (Patterson et al., 2007; Machery, 2016). ATL degrades in patients with SD, which correlates with semantic memory decline; this decline is not concept-selective (different categories of concepts decline in SD); ATL is not a part of the sensory-motor system; therefore, ATL may be the core place where concepts from different categories are stored in the amodal way.

Answering the argument regarding SD patients, it is true that ATL is indeed an important part of the semantic system, however, it is not the core area where concepts are stored in an amodal way. ATL is just a part of the concepts’ distributed circuits and, what is more important, is a connector hub for other areas which are also

included in the concept circuits (Kiefer & Pulvermüller, 2012; Garagnani & Pulvermüller, 2016). Moreover, ATL is not a unique connector hub, as there are other areas that can play connector roles (Binder & Desai, 2011). In this study, we demonstrate exactly this idea by means of the biologically-inspired neuronal model and show how the SD mechanism leads to semantic deterioration, highlighting the role of ATL in this process.

The second argument states that there exists an asymmetry between activation of the sensory-motor system during concept processing and problems with concept comprehension in patients who have a lesion in a particular sensory or motor area. (Mahon & Hickok, 2016; for review see Barsalou, 2016). It is a well-known result that parts of the sensory-motor system are active when we process corresponding concepts (ex. motor areas for action-related concept "grasp" or visual areas for object-related concept “apple”) (for review see Kiefer & Pulvermüller, 2012; Meteyard et al., 2012). However, there still exist patients who have lesions the sensory-motor system but have little or no deterioration of corresponding conceptual knowledge — for instance, a patient cannot move legs anymore, but she can understand the concept of running (Mahon & Hickok, 2016). Therefore, this asymmetry shows that the sensory-motor system plays only a secondary role in concept representations — we do not really need it for concept comprehension and its activation during concept processing is just a by-product of the amodal representations’ activation.

Answering this argument, we, first of all, should note that there are many experiments which show that patients with sensory and motor lesions have deficits in the corresponding conceptual knowledge (for review see Kiefer & Pulvermüller, 2012; Meteyard et al., 2012). However, the existence of patients who have little or no deficits does not contradict hybrid theories. Concept representations are not restricted to primary sensory or motor areas, but are distributed through other areas as well: secondary sensory-motor areas, connector hubs, language areas, and so on (Kiefer & Pulvermüller, 2012; Garagnani & Pulvermüller, 2016; Barsalou, 2016). Moreover, as we show in this study by means of a neuronal model, concept circuits are denser

(have more cells) in the central (connector) areas than in the primary ones, which means that problems in primary areas lead to fewer concepts deterioration compared with problems in more central areas. Therefore, a lesion in the primary sensory or motor area can disrupt only a small part of the concept circuits, which means that we have to create very sensitive tests to catch these small disruptions (Barsalou, 2016). This can explain the existence of some patients with lesions in the sensory-motor system who have little or no concept knowledge deterioration.

The third argument states that, if congenitally blind people who have no first-hand visual experience can form visually-related concepts that are very similar to the sighted population's concepts, then conceptual knowledge can be easily acquired without sensory experience (Mahon & Hickok, 2016; Leshinskaya & Caramazza, 2016). Therefore, the sensory-motor system plays only a secondary role in the semantic system.

Answering the argument regarding congenitally blind people, first of all, we should again note that there are papers that show differences in concept knowledge between congenitally blind and sighted people. For instance, experiments with one of the most visually-related concepts — colors — show that congenitally blind individuals can operate these concepts but much worse than sighted individuals (Shepard & Cooper, 1992; Saysani, Corballis, & Corballis, 2018). Researchers suggest that congenitally blind people can partially learn relationships between colors through colors' names co-occurrence in language.

Secondly, even in papers which are usually used as proof of semantic knowledge similarities between two populations, there exist differences between these populations, though they tend to be discussed briefly or are omitted from the discussion completely. For instance, Bedny, Koster-Hale, Elli, Yazzolino, and Saxe (2019) report that blind people have judgments about visual verbs very similar to those of sighted people. However, at the same time, they report that blind people have more coherence among themselves in judgments about touch perception and sound emission verbs than sighted people do, thus showing the difference between these populations. Mahon, Anzellotti, Schwarzbach, Zampini, and Caramazza (2009) show

that for blind and sighted individuals the same areas distinguish two types of concepts: medial fusiform gyrus for nonliving things and lateral occipital cortex for animals. However, as we can see from the reported pictures, patterns of activation actually differ for two populations. For instance, in sighted people lateral occipital cortex is more active for animals while in blind people some parts of this ROI and areas next to it are at the same time more active for nonliving things (see Mahon et al., 2009, Figure 3A). In another paper, authors answer a different question from the one they are asking (Bedny, Caramazza, Pascual-Leone, & Saxe, 2012). They show that the same area (lMTG) in blind and sighted people responds more strongly to verbs than to nouns. But they fail to show that lMTG responds to motion properties for both verbs and nouns in both groups of people (which is what they initially want to show, as they use motion degree as a visual feature). Using these results, they cannot conclude that activation of lMTG reflects semantics, at least not the visual features of semantics. Therefore, activation of lMTG — which is similar for congenitally blind and sighted populations — cannot be used as a proof that blind people can develop undistinguishable from sighted people visually-related concepts without visual experience.

What is more important, hybrid theories assume that concepts are acquired through co-experience of many different sensory-motor and language inputs (Kiefer & Pulvermüller, 2012; Barsalou, 2016). For example, mom gives an apple to her child, the child hears the word ‘apple’, sees an apple, touches its surface, tastes it, and so on. Therefore, even if a congenitally blind person has no access to first-hand visual information, inputs from other sensory modalities or learning through words co-occurrence in language can partially compensate for the problem of concept acquisition. This, in turn, can lead to activations of the same areas as in sighted people. Therefore, it is not enough to show that the same areas are active in two populations, as done in many papers (Bedny et al., 2012; Mahon et al., 2009). Instead we have to track the detailed profile of activation — a spread of activity between areas, an amplitude, a shape of the BOLD signal, and so on — to catch differences. In this paper, we use a neuronal model to show how the compensation mechanism

discussed above can influence concept representations in the blind and what predictions for future experiments can be made taking this into account.

2.2 Approaches to Semantic System Modeling

Here we describe different approaches to semantic system modeling and their pitfalls and limitations and then discuss why we believe that the model we work on (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018) is one of the most biologically precise models that exist to date even despite its limitations.

The present-day approaches to modeling of the semantic system consist of the two extremes and the niche for further research in between. On the one extreme are purely statistical models which correlate meanings of real words with brain activations across the whole cortex (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Anderson et al., 2016). On the another extreme are bio-inspired neurocomputational models, which are presented by a limited number of the cortex areas and only get abstract synthetic data as input (Chen et al., 2017; Garagnani & Pulvermüller, 2016; Tomasello et al., 2018).

Huth et al. (2016) try to create the map of semantic representations in the cortex. In the experiment, seven participants listened to a story for more than two hours in the fMRI scanner. Each word in the story was projected onto a 985-feature space which consists of 985 popular English words. The projection is based on the similarity of context between the word in the story and each of 985 popular English words (e.g. “giraffe” and “zebra” are often used in a similar context while “giraffe” and “socket” are not). BOLD signal in each voxel of the cortex is regressed on 985 features and controls, which helps to catch only semantic component of words and avoid others, for example, phonetic ones. To reveal which semantics activate each voxel — therefore, to create a semantic map — authors further perform principal component analysis (PCA) and cluster all words. They get 12 categories which they label manually in the following way: tactile, visual, numeric, locational, abstract, temporal, professional, violent, communal, mental, emotional. Then, they create a semantic map of the cortex in accordance with these categories and show that this map is more or less consistent across different individuals. A very similar approach to

create a semantic map is proposed by Anderson et al. (2016). They also use fMRI recording (in this case, during the sentence reading) and apply a purely statistical algorithm (multiple regression) to predict cortex activation in response to particular words. However, they use predefined in the previous study "semantic features" (Binder et al., 2016), which are distributed across different categories: visual, auditory, somatosensory, gustatory, somatosensory, motor, attentional, event, evaluation, cognitive, emotional, drive, spatial.

On the other extreme, Chen et al. (2017) suggest that the organization of semantic representations can be explained by a combination of three factors: the sensory-motor experience, brain connectivity, and associative learning. They implement these ideas in a model based on a biologically constrained deep recurrent neuronal network in the hub-and-spoke paradigm (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 of one amodal hub in the anterior temporal lobe (ATL). The choice of these areas is based on previous studies which revealed the most important brain regions for semantic processing, with connectivity between these areas being assessed via DTI. Further, the authors create two different semantics — animals and tools. It is suggested that animal comprehension relies mostly on visual experience, therefore, the model is taught to associate activation in the visual and auditory areas to create semantic representations of animals. For tools, on the other hand, motor experience is more important than the visual one, therefore, the model is 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 shows higher activation in the visual regions for animal-related input and higher activation in the function-praxis regions for tool-related 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 suggest that sensory-motor co-experience during concept acquisition, brain connectivity, and associative learning mechanism are the fundamental principles that can explain the empirical data about semantic representations. However, they do not assume predefined locations for semantic representations, as done by Chen et al. (2017). On the contrary, they try to simulate many different cortical regions to show how the semantic system can emerge on the basis of these regions under fundamental principles of the brain anatomy and functioning. This approach shows that both different modality-specific areas and different connector hubs are important for the semantic system, as assumed by hybrid theories. The model consists of four zones, with each zone containing three cortical areas: 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 decades-long research, as described in Garagnani & Pulvermüller (2016). Another important distinction of this model is biologically constrained fine architecture of the neuronal network and biologically constrained learning algorithm. Each area consists of two layers of neurons — excitatory and inhibitory ones. Excitatory neurons are modeled by graded-response neurons in the early model iteration (Garagnani & Pulvermüller, 2016) and by spiking neurons in the latest one (Tomasello et al., 2018; Tomasello, Wennekers, Garagnani, & Pulvermüller, 2019); inhibitory neurons are modeled by graded-response neurons. Connections between and within areas are sparse, topographically constrained and random (with decreasing probability, which depends on the distance between two neurons). Notably, the

authors use the Hebbian learning mechanism ("fire together — wire together") — an associative learning mechanism that is assumed to be used in the brain. All neurobiological features which we discuss above significantly distinguish this model from the model proposed by Chen et al. (2017), where more mechanical recurrent neural network, predefined locations for semantic representations and backpropagation as the learning mechanism are used. The problem of such modeling (Chen et al., 2017) is that this architecture and learning mechanism do not try to replicate the real properties of the brain. Therefore, such a model can only fit the empirical data but cannot provide an explanation of the processes in which this data appears.

Authors (Garagnani & Pulvermüller, 2016) create two different semantics: object- and action-related words. The model is 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 is 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, cell assemblies are assessed by presenting only auditory and articulatory patterns. It is shown that these patterns for object-related words propagate up to the primary visual area when almost no activity is registered in the primary motor area. The propagation dynamics is opposite for the motor-related words — it reaches the primary motor area but not the primary visual one. This result means that the model learns to discriminate between different categories of semantics on the basis of sensory-motor representations, which is consistent with the empirical data (for review see Kiefer & Pulvermüller, 2012; Barsalou, 2016). Notably, the central areas show almost no distinction between two categories of semantics, therefore, they are mostly category-general. This is consistent with the empirical findings that there exist several multimodal connector hubs, while still having modality-specific areas play a key role in concept representations (Binder & Desai, 2011; Kiefer & Pulvermüller, 2012).

In this study, we are going to use the model proposed by Pulvermüller and colleagues (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018) as the basis to which we add the semantic dementia mechanism. We describe this model in more detail in section *3.1 Baseline Model Description*.

2.3 Semantic Dementia

Semantic dementia (SD) is a progressive neurological disorder which is characterized by semantic knowledge deterioration. The degree of this deterioration correlates 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, Saito, Rogers, & Lambon Ralph, 2011; Chen et al., 2017) and TMS studies (Pobric, Jefferies, & Lambon Ralph, 2007), the ATL atrophy is considered as the cause of problems with semantic memory. Therefore, SD is a key disease to consider if we want to explore how the semantic system works and what neuronal mechanisms underlie it.

The main symptoms of SD are anomia and difficulty with word comprehension, while word repetition abilities remain better preserved (for review see Hodges & Patterson, 2007; Montembeault et al., 2018). Anomia is difficulty with retrieval and usage of right words, for instance, when an SD patient tries to describe something or to tell a story. When an SD patient has difficulties with word comprehension, she misunderstands some words or the whole speech. Although SD patient cannot understand the word she can repeat it — for instance, Hodges, Martinos, Woollams, Patterson, and Adlam (2008) shows that patients cannot point to the picture of ‘hippopotamus’ (which suggests that they do not know the meaning of this word) but repeat this complex word easily. Moreover, such people have difficulties with understanding when they have to interact with concepts in non-verbal form. For example, an 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

stages 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) and from their works we know about the typical errors SD patients make (Hodges et al., 1995). SD patients sometimes call the object of interest as a different object in the same category — for example, “zebra” for “giraffe”. Or they call the object of interest by its superordinate category — for example, “animal” for “zebra”. Or they describe the object of interest instead of naming it — for example, “horse in a desert” for “camel”. Or they use a semantic association instead of the word itself (“tree” for “forest”). Or they can call the object of interest as some completely unrelated object (“apple” for “bird”).

The dynamic pattern of semantic deterioration is highly robust: it is a progressive loss that starts with the degradation of specific semantic details and spreads to more general ones. First, differences between specific close categories disappear: for example "zebra" starts to get confused for “giraffe". Then, the superordinate category can 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 concepts, as well as more atypical concepts, are deteriorated earlier during the disease course (Hodges & Patterson, 2007; Montembeault et al., 2018; Rogers, Patterson, Jefferies, & Lambon Ralph, 2015).

Although some researchers suggest that semantic deterioration during SD is category-general, meaning that all categories are deteriorated almost in the same manner and at the same disease stage (Patterson et al., 2007; Lambon Ralph, Lowe, & Rogers, 2007), we found no experiments that directly compare the degradation dynamics of object-related words with the degradation dynamics of action-related words (due to our model specification we are interested in these two categories).

Usually, nouns and verbs are compared in such cases (Daniele, Giustolisi, Silveri, Colosimo, & Gainotti, 1994; Breedin, Saffran, & Coslett, 1994; Yi, Moore, & Grossman, 2007). However, they are a poor proxy for semantic categories (object- versus action-related words) as the distinction between nouns and verbs includes confounding syntactic information (Pulvermüller et al., 2010). Moreover, some verbs depend heavily on visual features — like the verb "blink" — while nouns can depend on motor features — like the noun "run" — which should be controlled carefully. In addition, on average, nouns are used less frequently than verbs, which also can influence SD degradation dynamics (Bird, Lambon Ralph, Patterson, & Hodges, 2000). Even controlling for the frequency factor, we still have contradictory results in the literature: some researchers find that nouns degrade more heavily than verbs (Daniele et al., 1994), others find the opposite pattern (Yi et al., 2007). We suggest that object- versus action-related categories should be considered within verbs or within nouns separately. In such experiments, it is found that nouns and verbs that are tightly related to visual features degrade more dramatically than other nouns and verbs. For instance, forms of objects remain better-preserved than colors which are more visually-related concepts (Pulvermüller et al., 2010), as forms of objects have a tactile dimension in addition to the visual dimension. Abstract verbs and nouns are better preserved than concrete verbs and nouns (Bonner et al., 2009; Yi et al., 2007; Cardebat, Demonet, Celsis, & Puel, 1996), as latter have more visual features — it is not easy to create a picture of the concept "friendship", but we can easily do this with the concept "apple". To sum up, from the existing literature we can get only some clues that for SD patients object-related concepts, which, by definition, are tightly related to the visual features, should be damaged more dramatically than the action-related concepts. However, we suggest that more explicit experiments are needed to understand this dynamics better.

Neuronal anatomical and functional changes during SD can be described as a bilateral and asymmetrical pattern of the progressive atrophy and hypometabolism in white and gray matter (Hodges & Patterson, 2007; Montembeault et al., 2018; Spinelli et al., 2017). Gray matter changes are detected in the anterior temporal lobes

— superior, middle, inferior temporal gyri, fusiform gyrus, temporal pole, parahippocampal 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 regions that are connected with or are adjacent to 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 most severe and the most robust across the SD patients atrophy is detected in the anterior temporal pole and ventral parts of the ATL (Lambon Ralph et al., 2017).

Degradation is often bilateral and asymmetrical, and usually more severe in the left ATL (Hodges et al., 2009). However, asymmetrically stronger degradation in the right ATL occurs as well. The current consensus is that atrophy in the left ALT mostly leads to anomia and to difficulties with word 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).

In this study, we build an SD model and test our hypothesis about the main pattern of the SD dynamics, namely that word recognition abilities decline with an increase in the disease severity, while word repetition abilities remain better preserved. We implement two types of SD damage — gray matter and white matter damage of the ATL — and compare model's results for them. We also compare the degradation dynamics of object-related words with the degradation dynamics of action-related words during SD progression.

Chapter 3. Methods

The model proposed by Pulvermüller and colleagues (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018) is one of the most biologically accurate models of the semantic system existing today (for review see section *2.2 Approaches to Semantic System Modeling*). For simplicity, we use the early iteration of this model (Garagnani & Pulvermüller, 2016) to implement the SD mechanism. Although the latest model iteration is more complex and more biologically realistic (it includes improved connectivity structure and spiking neurons) the early and the latest iterations give qualitatively similar results.

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, one amodal semantic hub, and less biologically constrained architecture (e.g. Chen et al. (2017) use an artificial recurrent neural network with backpropagation learning mechanism). Thus, it appears that previous models, even if fitting the data well, do not explain the biological mechanisms of semantic representations. Therefore, we believe that the current study can contribute to the better understanding of the semantic system in the brain.

In this chapter we describe the full specification of the baseline model created by Garagnani & Pulvermüller (2016), as we replicate their results, prior to implementation of SD into their model. Then, we discuss the specification of the SD mechanism that we create and formulate the predictions to test our model.

3.1 Baseline Model Description

3.1.1 Macro-structure. The original model consists of four modality-specific zones, with each zone containing three cortical areas (see Figure 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). We differentiate between the perisylvian cortex (auditory and articulatory areas) and the extrasylvian cortex (visual and motor areas). We also differentiate between the frontal cortex (articulatory and motor areas) and the temporal cortex (auditory and visual areas).

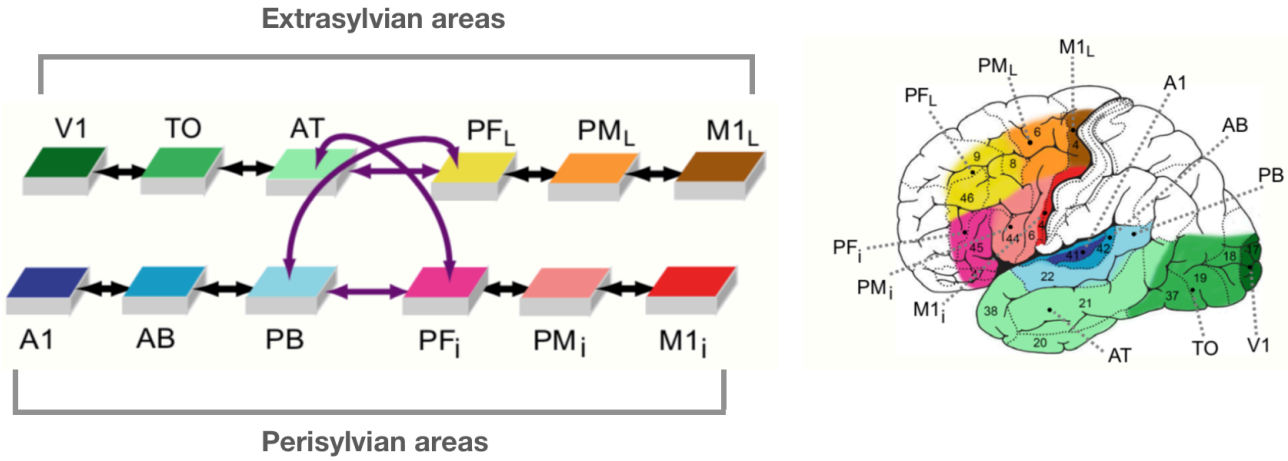

Figure 1. Four cortex zones — auditory (blue colors), articulatory (pink-red colors), visual (green colors) and motor (yellow-brown colors) — with 3 areas within each zone and a specific connectivity pattern (black arrows indicate connections between two adjacent areas; purple arrows indicate long-distance cortico-cortical connections). Adapted from the *Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs* (Figure 1 A, B), by Garagnani, M., & Pulvermüller, F., 2016, European Journal of Neuroscience, 43(6) Copyright 2015 The Authors. European Journal of Neuroscience published by Federation of European Neuroscience Societies and John Wiley & Sons Ltd.

This model extends a previous model in which only six areas of the traditional language (perisylvian) cortex were modeled — auditory and articulatory zones

(Garagnani, Wennekers, & Pulvermüller, 2007). Six additional areas in the visual and motor zones (extrasylvian cortex) allow to build the basis for the semantic system.

Two types of cortico-cortical connections are suggested in the model. The first type is connections between two adjacent areas (depicted by black arrows in Figure 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. The second type is long-distance cortico-cortical connections (depicted by purple arrows in Figure 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 decision to include these connections is based on decades-long research on the neuroanatomical structure of the cortex (for review see Garagnani & Pulvermüller, 2016).

3.1.2 Micro-structure. Each area consists of two neuronal layers — an excitatory (e-cell) one and an inhibitory (i-cell) one — with 625 (25x25) cells in each layer (see Figure 2). Each i-cell corresponds to exactly one e-cell; a combination of an e-cell and an i-cell reflects approximately one cortical column that consists of pyramidal excitatory neurons and inhibitory interneurons. Excitatory and inhibitory neurons are modeled as graded-response neurons.

To understand fine connectivity within and between areas, let us consider one of the e-cells (see Figure 2).

First, we will explain the interaction of this e-cell with other e-cells. The e-cell sends its projections to the 19x19 e-cell patch in the same area, the e-cell under consideration being the central one in this patch. It also sends its projections to a topographically similar 19x19 e-cell patch in another area (with which the current area is connected through adjacent or long-distance cortico-cortical connection, as discussed in subsection *3.1.1 Macro-structure*). These projections to the same-area or to the different-area patches are created in a random manner, with probability that the projection is created being subject to Gaussian density function centered in the patch

center[1]; link weights are initialized at random with uniform distribution[2]. Likewise, other e-cells send projections to the e-cell of interest.

Second, we will consider the interaction of this e-cell with i-cells in the same area. Each i-cell corresponds to exactly one e-cell and sends its projection only to this one e-cell. At the same time, this i-cell gets projections from each e-cell in the 5x5 patch around the e-cell of interest. The i-cell sums up inputs from the patch of e-cells and inhibits the e-cell of interest proportionally to this sum, reflecting the local inhibition mechanism. In this way, each e-cell sends projections to the 5x5 patch of i-cells to inhibit its neighboring e-cells.

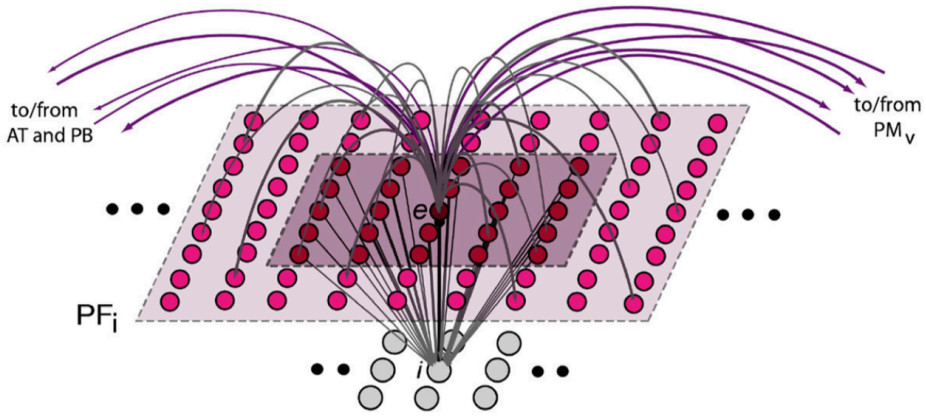

Figure 2. Micro-structure of the model, an example from the PF_i area. Pink area reflects the e-cells layer with the e-cell of interest in the middle of the patch, grey units are i-cells with the i-cell in the middle that corresponds to the e-cell of interest. Reprinted from the C*onceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs* (Figure 1 C), by Garagnani, M., & Pulvermüller, F., 2016, European Journal of Neuroscience, 43(6) Copyright 2015 The Authors. European Journal of Neuroscience published by Federation of European Neuroscience Societies and John Wiley & Sons Ltd.

[1] See parameters in Appendix

[2] See parameters in Appendix

3.1.3 Dynamics of the model. First, the dynamics of e-cells will be considered (see Figure 3)[3].

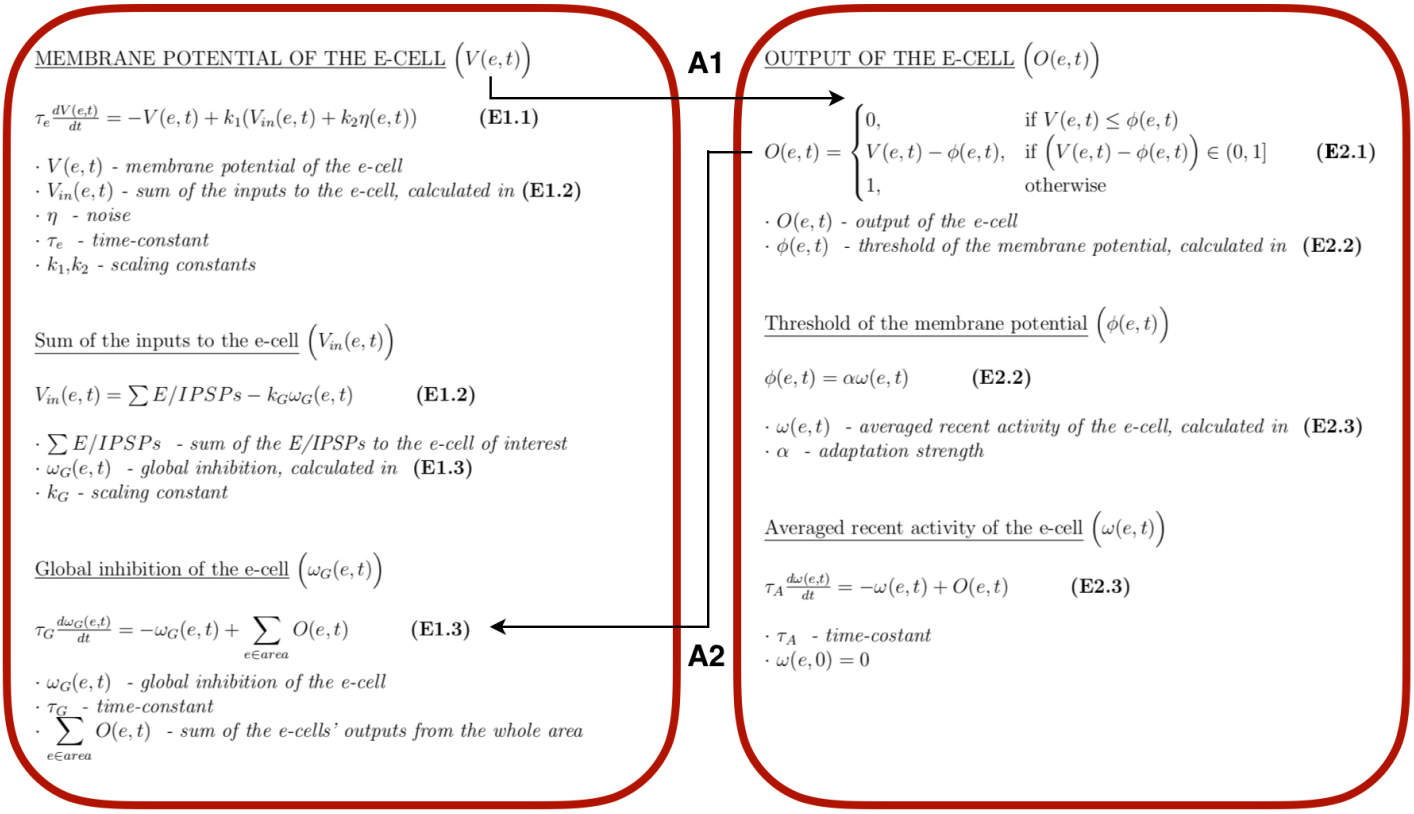

Figure 3. Dynamics of the e-cell: the membrane potential of the e-cell (left panel) and the output of the e-cell (right panel). Arrows indicate that **(A1)** the membrane potential is used for the cell output calculation and **(A2)** the cell output is used for the global inhibition calculation which is later used in the membrane potential calculation.

Changes in the membrane potential of the e-cell depend on the current membrane potential, the sum of the inputs to this e-cell, and the white noise specific to e-cells (see **(E1.1)**). The sum of inputs to the e-cell consists of the sum of all excitatory and inhibitory postsynaptic potentials (E/IPSPs) and global inhibition component (see **(E1.2)** and **(E1.3)**). EPSP from e-cell *x* to e-cell *y* (as well as IPSP from i-cell *x* to e-cell *y*) is defined as the output of cell *x* multiplied by the weight of

[3] See parameters in Appendix

the synaptic connection from cell *x* to cell *y* (weights initialization and the rule according to which weights are changed are explained further). Each e-cell gets only one IPSP from the corresponding i-cell, which is included into this sum with a negative sign. White noise is an important biological feature of this model. It is only present in e-cells and it simulates spontaneous activity of excitatory neurons. The e-cell output depends on the value of its membrane potential (depicted by the arrow **A1**) and the threshold (see **(E2.1)**), 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 **(E2.2)** and **(E2.3)**). Global inhibition of e-cells is an important mechanism in this model. It is calculated using the sum of e-cells' output from the same area as the e-cell under consideration (see **(E1.3)**) — the higher activity of all other e-cells in the area is the stronger inhibition of the e-cell of interest will be. This global inhibition influences the net input to the e-cell, which in turn shapes the membrane potential (see **(E1.2)** and **(E1.1)**).

Second, the dynamics of i-cells will be considered (see Figure 4)[4]. In this case, changes in membrane potential of the i-cell depend on the current membrane potential and sum of inputs to this i-cell (see **(I1.1)**). The sum of inputs to the i-cell consists only of EPSPs from the 5x5 e-cells patch (see **(I1.2)**). I-cell output depends only on the value of its membrane potential (depicted by the arrow **A3**): if membrane potential is greater than zero, inhibition occurs (see **(I2.1)**).

Finally, we will discuss initialization of synaptic weights and the rule according to which they are changing during learning[5]. Initially, random weights are assigned to all established connections (recall that connections are also created randomly, subject to the Gaussian density function). Then, Hebbian learning takes place (see Figure 5).

[4] See parameters in Appendix

[5] See parameters in Appendix

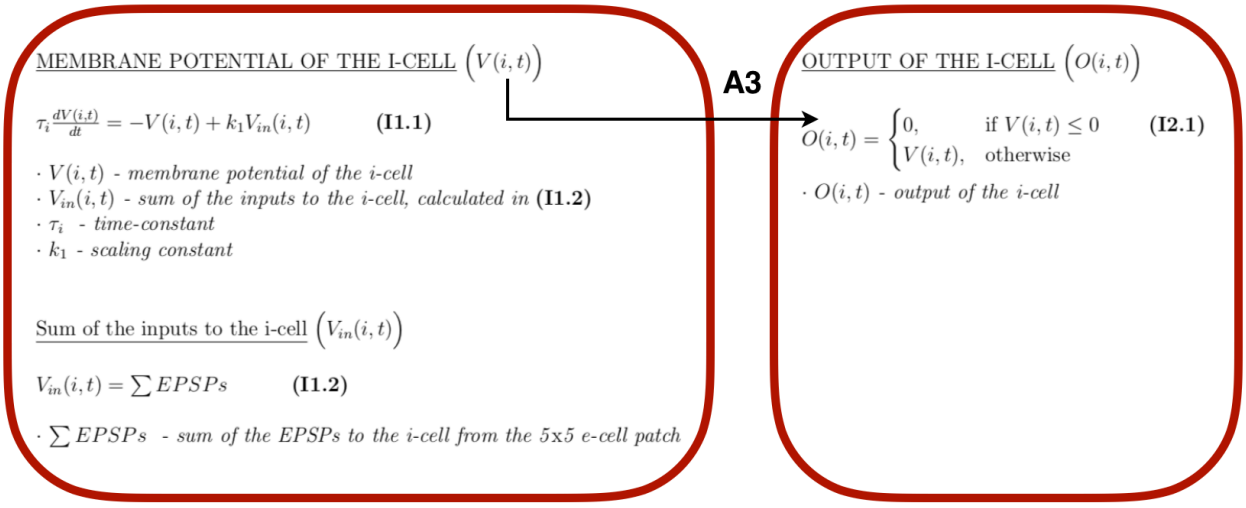

Figure 4. Dynamics of the i-cell: the membrane potential of the i-cell (left panel) and the output of the i-cell (right panel). Arrow **(A3)** indicates that the membrane potential is used for the cell output calculation.

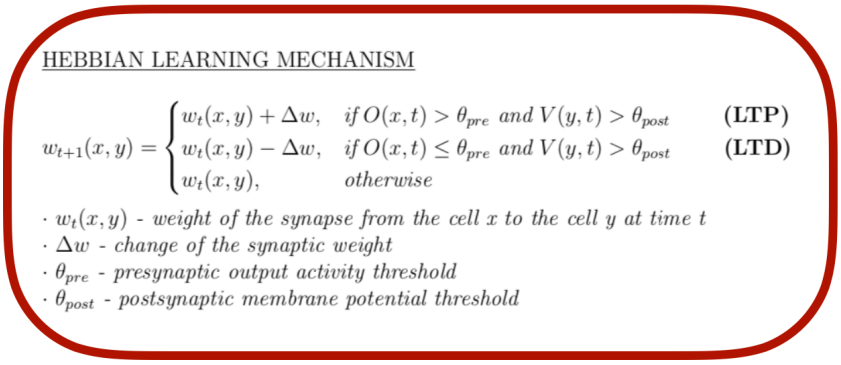

Figure 5. Hebbian learning mechanism. Long-term potentiation (**LTP**) and long-term depression (**LTD**) equations.

If high enough pre-synaptic activity (cell's output of the pre-synaptic cell) corresponds to high enough post-synaptic activity (membrane potential of the post-synaptic cell), synaptic weight is strengthened (long-term potentiation — **(LTP)**), subject to the rule "fire together — wire together". If low pre-synaptic activity corresponds to high post-synaptic activity, synaptic weight is weakened (long-term depression — **(LTD)**), subject to the rule "out of sync — out of link" (Artola & Singer, 1993).

3.1.4 Semantic categories. The baseline model is taught to differentiate between two semantic categories: action- and object-related words. The model acquires object-related semantics by co-experiencing and associating auditory, articulatory and visual patterns, as inputs to A1, $M1_i$ and V1 are presented simultaneously during the learning phase (see Figure 6A). To teach action-related semantics, inputs are provided to auditory (A1), articulatory ($M1_i$) and motor ($M1_L$) primary areas (see Figure 6B).

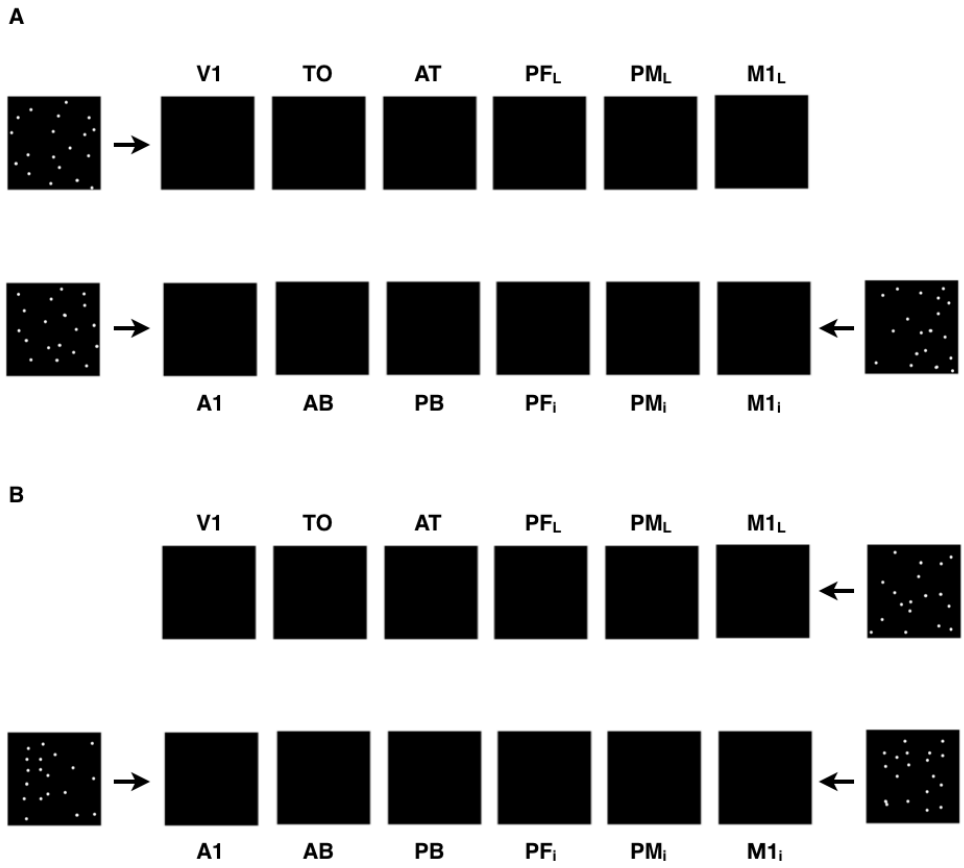

Figure 6. An example of object-related word acquisition via inputs to auditory (A1), articulatory ($M1_i$) and visual (V1) primary areas (**A**); an example of action-related word acquisition via inputs to auditory (A1), articulatory ($M1_i$) and motor ($M1_L$) primary areas (**B**).

Six word patterns for each of the two semantic categories are generated, therefore, the model learns 12 different word patterns. This mechanism of co-experiencing of different perceptual inputs to learn semantics reflects the idea of

teaching a child. For example, mom teaches her child to name “cat”: she pronounces the word “cat”, asks her child to repeat, and at the same time points to the cat.

3.1.5 Learning and adjustment procedures. Thirteen nets with the architecture described above are initialized randomly and pass through the following learning and adjustment procedures.

During the learning procedure, each net is taught a word pattern by simultaneous presentation of inputs to three relevant modalities out of four ones, as described in the previous subsection *3.1.4 Semantic categories*. Each word pattern is presented for 3000 trials, therefore, the learning phase lasts 36000 trials (3000x12). One trial lasts 16 time-steps. The next trial starts as soon as the activity in the network falls below the threshold but not earlier than 30 time-steps from the previous trial’s end. Trials are randomly shuffled.

It is important to note that a random 19-cell noise pattern is provided to the non-involved primary area (V1 for action-related words and $M1_L$ for object-related words). This noise pattern changes in each trial, so it differs from the stable 19-cell patterns to the primary areas which are involved in semantic acquisition. In addition, white noise is presented to all four primary areas, as it reflects spontaneous neuronal activity in the areas.

After the learning phase, cell assemblies (CAs) for all 12 words emerged (see Figure 7), only excitatory cells are included into the CAs. To evaluate these CAs the net is passing through adjustment procedure. Each word pattern is presented only once (for 16 time-steps) to the language areas — auditory (A1) and articulatory ($M1_i$) ones, while nothing is presented to the semantic areas — visual (V1) and motor (M1) ones. After this presentation, the activity of each e-cell is recorded for the next 15 time-steps. E-cell is added to the CA if its time-averaged activity reached a threshold (θ_{adj}).

The threshold is specified for each word (w) and each area (a) separately in the following way:

$$\theta_{adj}(w,a) = \gamma \max_{e \in a} \overline{O(e,t)_w}$$

Meaning that the time-averaged output is calculated for each e-cell (*e*) in each area (*a*) per each word (*w*). The threshold for the area for the word is equal to the fraction (γ) of time-averaged output of the most active cell in this area for this word.

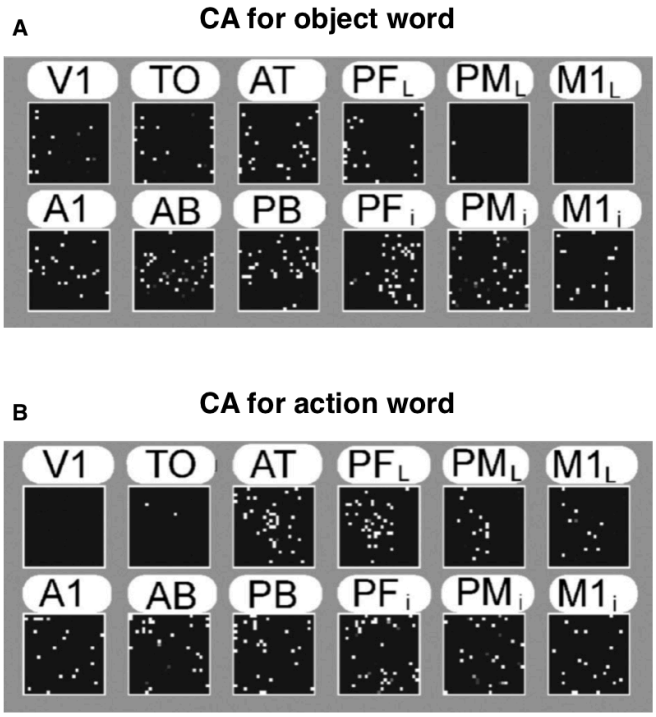

Figure 7. Examples of an object-related CA (**A**) and an action-related CA (**B**). Adapted from the *Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs* (Figure 2), by Garagnani, M., & Pulvermüller, F., 2016, European Journal of Neuroscience, 43(6) Copyright 2015 The Authors. European Journal of Neuroscience published by Federation of European Neuroscience Societies and John Wiley & Sons Ltd.

3.2 Semantic Dementia Implementation

We create two types of degradation to simulate SD — grey matter (GM SD) and white matter (WM SD) degradation, as we know from patients' neuroimaging data that both types of damage take place (Hodges & Patterson, 2007; Montembeault

et al., 2018; Spinelli et al., 2017). To simulate GM damage we turn off e-cells in the AT area (see Figure 8A) and to simulate WM damage we remove to-, from-, and recurrent links of the AT area (see Figure 8B).

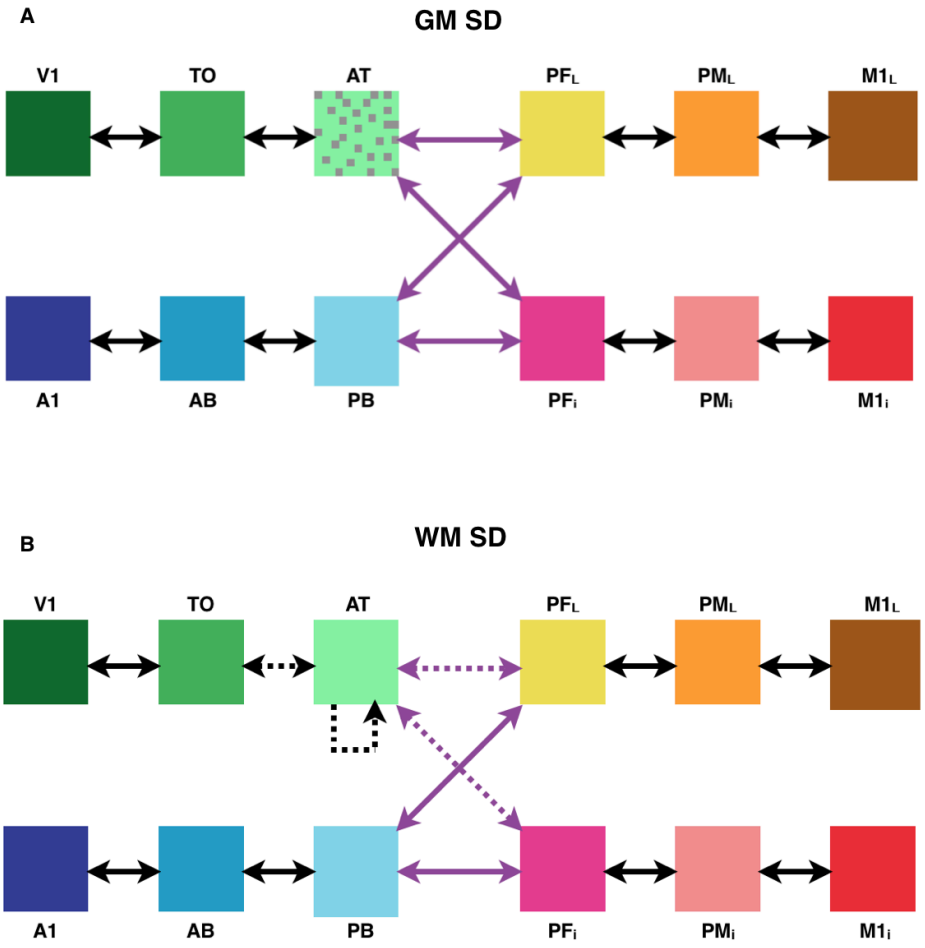

Figure 8. Illustration of gray matter semantic dementia (GM SD) — degradation of cells in the AT area (**A**) — and white matter semantic dementia (WM SD) — degradation of to-, from- and recurrent links in the AT area (**B**).

To the best of our knowledge, this is the first model where both degradation types were implemented, previously only degradation of links (WM) was used (Ueno et al., 2011; Chen et al., 2017). Thus, we compare the impact of these two types of damage on model's recognition abilities. In future experiments, we can combine them to make better predictions about semantic system deterioration.

We also use three severity levels for each degradation type — 30%, 60%, and 90% loss of matter — to simulate effects of the progressive nature of this disorder (Hodges & Patterson, 2007; Montembeault et al., 2018; Spinelli et al., 2017).

We train 13 nets using the learning procedure described above and only then implement one of two degradation types (GM SD or WM SD) of the particular severity level (30%, 60%, or 90% loss of matter). Doing so, we replicate real situation with SD patients who acquired semantic knowledge before having any problems with ATL.

After SD implementation, we run a procedure to evaluate the recognition abilities of the nets. Each word pattern is presented once (for 16 time-steps) to the auditory area only (A1). After this presentation, the activity of each e-cell is recorded for the next 15 time-steps. E-cell is added to the cell circuit if its time-averaged activity reached a threshold (θ_{rec}). This threshold is specified for each word (w) and each area (a) separately in exactly the same way as it is specified for the adjustment procedure.

$$\theta_{rec}(w, a) = \gamma \max_{e \in a} \overline{O(e, t)_w}$$

3.3 Model's Predictions

As we noted in section *2.3 Semantic Dementia*, our hypothesis regarding SD progression is the following: word recognition abilities decline with SD severity increase, while word repetition abilities remain better preserved (as suggested by patients' data (Hodges & Patterson, 2007; Montembeault et al., 2018)).

To test this hypothesis in our model we make the following predictions: activated cell circuits after the recognition procedure decrease dramatically in extrasylvian areas with SD progression, while in perisylvian areas they remain mostly intact. We based these predictions on the previous results obtained from the original model (Garagnani & Pulvermüller, 2016; Garagnani et al., 2007) which show that extrasylvian areas are primarily responsible for semantic knowledge (therefore, we assume that they contain the recognition part of concept circuits), while perisylvian

areas are primarily responsible for language circuits (therefore, we assume that they contain the repetition part of concept circuits).

Chapter 4. Results

4.1 Replication of the Baseline Model

We repeat experiments from Garagnani & Pulvermüller (2016) to ensure that our baseline model (to which we add our SD mechanism) works properly in the new environment. For this replication, we use model specification described in section *3.1 Baseline Model Description* and measure the number of CA cells in each area for each word, as described in subsection *3.1.5 Learning and adjustment procedures.*

Visual inspection of the results (see Figure 9) confirms findings from Garagnani and Pulvermüller (2016). There is double dissociation between object- and action-related words in extrasylvian areas — more CA cells in primary and secondary visual areas (V1 and TO) for object-related words and more CA cells in primary and secondary motor areas ($M1_L$ and PM_L) for action-related words. In contrast, perisylvian areas do not differ dramatically in response to both types of words. To verify these observations, we run statistical tests that were used by Garagnani and Pulvermüller (2016).

4-way RM ANOVA with factors ExtraPeri (two levels: extra = {V1, TO, AT, PF_L, PM_L, $M1_L$} and peri = {A1, AB, PB, PF_i, PM_i, $M1_i$}), FrontoTemp (two levels: frontal = {$M1_L$, $M1_i$, PM_L, PM_i, PF_L, PF_i} and temporal = {V1, A1, TO, AB, AT, PB}), Centrality (three levels: primary = {V1, $M1_L$, A1, $M1_i$}, secondary = {TO, PM_L, AB, PM_i}, and central = {AT, PF_L, PM_L, PM_i}) and WordType (two levels — object and action) shows significant interaction of all four factors ($F_{2,24} = 179$, $p < 0.001$) and main effects of Centrality ($F_{2,24} = 1992.74$, $p < 0.001$) and ExtraPeri ($F_{1,12} = 99.34$, $p < 0.001$)[6], in accordance with Garagnani and Pulvermüller (2016). We perform dependent samples t-tests to check for Centrality effect and find that there are more CA cells in secondary areas than in primary areas ($t_{12} = 19.34$, $p < 0.001$) and more CA cells in central areas than in secondary areas ($t_{12} = 39.18$, $p < 0.001$), also in accordance with Garagnani and Pulvermüller (2016).

[6] All p-values for ANOVAs here and further are corrected for sphericity with the Huynh–Feldt correction where needed (more than two levels for the variable)

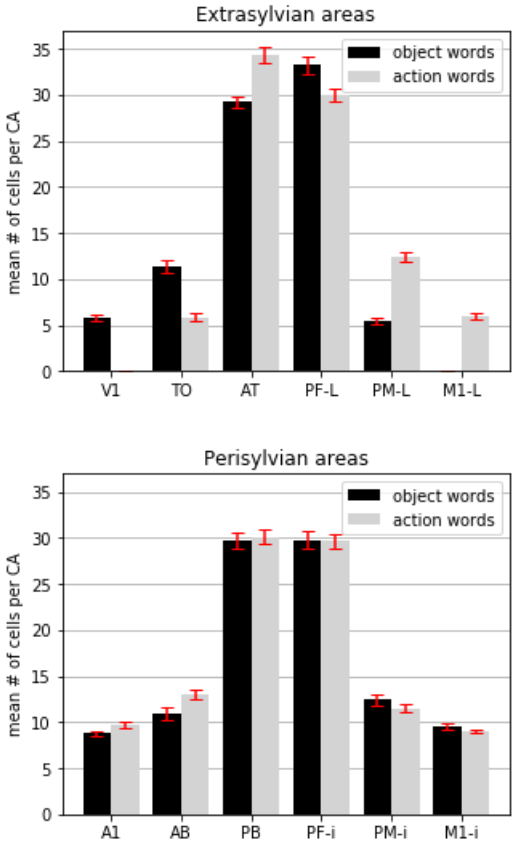

Figure 9. Replication of the baseline model's results (Garagnani & Pulvermüller, 2016). CA structure in extrasylvian (upper) and perisylvian (lower) areas for two word types: object-related words (black) and action-related words (grey); averaged across 13 nets. On the ordinate axis: the mean number of CA cells. Error bars indicate standard error of the mean.

Then, we separate extrasylvian and perisylvian areas and perform two 3-way RM ANOVAs with factors FrontoTemp, Centrality, and WordType. ANOVA for extrasylvian areas shows significant interaction of all three factors ($F_{2,24} = 311.74$, $p < 0.001$), while ANOVA for perisylvian areas does not ($F_{2,24} = 2.42$, $p = 0.11$), which confirm our suggestions that CA structure differs between extrasylvian and perisylvian areas. We also find the main effect of Centrality in both ANOVAs (for both ANOVAs: $F_{2,24} > 1315.22$, $p < 0.001$), also in accordance with Garagnani and Pulvermüller (2016). However, unlike Garagnani and Pulvermüller (2016), we find significant interaction of factors FrontoTemp and WordType in ANOVA for

perisylvian areas ($F_{1,12} = 14.45$, $p = 0.003$). Nevertheless, as we can see below, *post hoc* planned comparison tests (with Bonferroni correction, as in the original paper) do not show statistically significant differences between object- and action-related words in perisylvian areas.

We perform *post hoc* dependent samples t-tests to determine the word type effect on CA structure in each area separately and make Bonferroni correction for 12 comparisons (p-value threshold is 0.0042), as done by Garagnani and Pulvermüller (2016). As suggested from the visual inspection and found by Garagnani and Pulvermüller (2016), there are more CA cells for object-related than for action-related words in primary ($t_{12} = 16.27$, $p < 0.001$) and secondary ($t_{12} = 6.23$, $p < 0.001$) visual areas, while there are more CA cells for action-related than for object-related words in primary ($t_{12} = 18.27$, $p < 0.001$) and secondary ($t_{12} = 10.76$, $p < 0.001$) motor areas. However, unlike Garagnani and Pulvermüller (2016), we find a significant difference between two types of CAs in the central visual area — there are more CA cells for action-related than for object-related words ($t_{12} = 4.88$, $p < 0.001$). Nevertheless, this difference is much less prominent than differences between object- and action-related words in the primary or secondary visual areas. In accordance with Garagnani and Pulvermüller (2016), there is no significant difference in the central motor area ($t_{12} = 2.45$, $p = 0.03$). Also, there are no significant differences between object- and action-related words in all perisylvian areas (for all six areas: $t_{12} < 3.38$, $p > 0.006$).

These findings mostly replicate results from Garagnani and Pulvermüller (2016). We find double dissociation between object- and action-related words in extrasylvian areas and no such dissociation in perisylvian areas. This confirms our previous suggestion that extrasylvian areas are responsible for semantics, while perisylvian for language. We also confirm that different parts of the sensory-motor system play a crucial role for different concepts (here, object- versus action-related words). Moreover, in accordance with Garagnani and Pulvermüller (2016), we find that there exist semantic hubs in the central visual and motor areas, where substantial parts of CAs for both concept types are stored.

4.2 Semantic Dementia: Main Hypothesis Testing

Our predictions regarding SD progression are the following: number of cells in activated (by recognition procedure) circuits decreases dramatically in extrasylvian areas during SD progression, while in perisylvian areas it remains mostly intact.

We use model specification described in section *3.2 Semantic Dementia Implementation* and measure the number of cells in activated circuits after recognition procedure for each area per each word, as described there.

From the visual inspection of the results (see Figure 10), we can note that the number of cells in activated circuits in extrasylvian areas decreases dramatically due to the SD severity increase, while changes in perisylvian areas are much weaker. However, this dynamics is not identical within different extrasylvian areas; likewise it is not identical within different perisylvian areas. To explore this dynamics we proceed with statistical testing.

6-way RM ANOVA with factors ExtraPeri (two levels — extra and peri), FrontoTemp (two levels — frontal and temporal), Centrality (three levels — primary, secondary and central), WordType (two levels — object and action), SDType (two levels — GM and WM) and Severity (four levels — no SD, 30% SD, 60% SD, and 90% SD) shows significant interaction of all six factors ($F_{6,72} = 7.96$, $p < 0.001$). We also perform 2-way RM ANOVA with only two factors ExtraPeri and Severity, which are the most important factors for testing our hypothesis. It shows significant interaction of two factors ($F_{3,36} = 1166.29$, $p < 0.001$). This confirms our observation that extrasylvian and perisylvian areas respond differently to SD progression.

Then, we calculate the total number of cells in all six extrasylvian areas together and the total number of cells in all six perisylvian areas together. We perform 3-way RM ANOVAs for extrasylvian and perisylvian areas on this data separately with factors WordType, SDType and Severity. ANOVA for extrasylvian areas shows significant interactions of the factors SDType and Severity ($F_{3,36} = 73.04$, $p < 0.001$), it also shows significant interaction of the factors WordType and Severity ($F_{3,36} = 6.46$, $p = 0.0052$). While ANOVA for perisylvian areas shows only significant

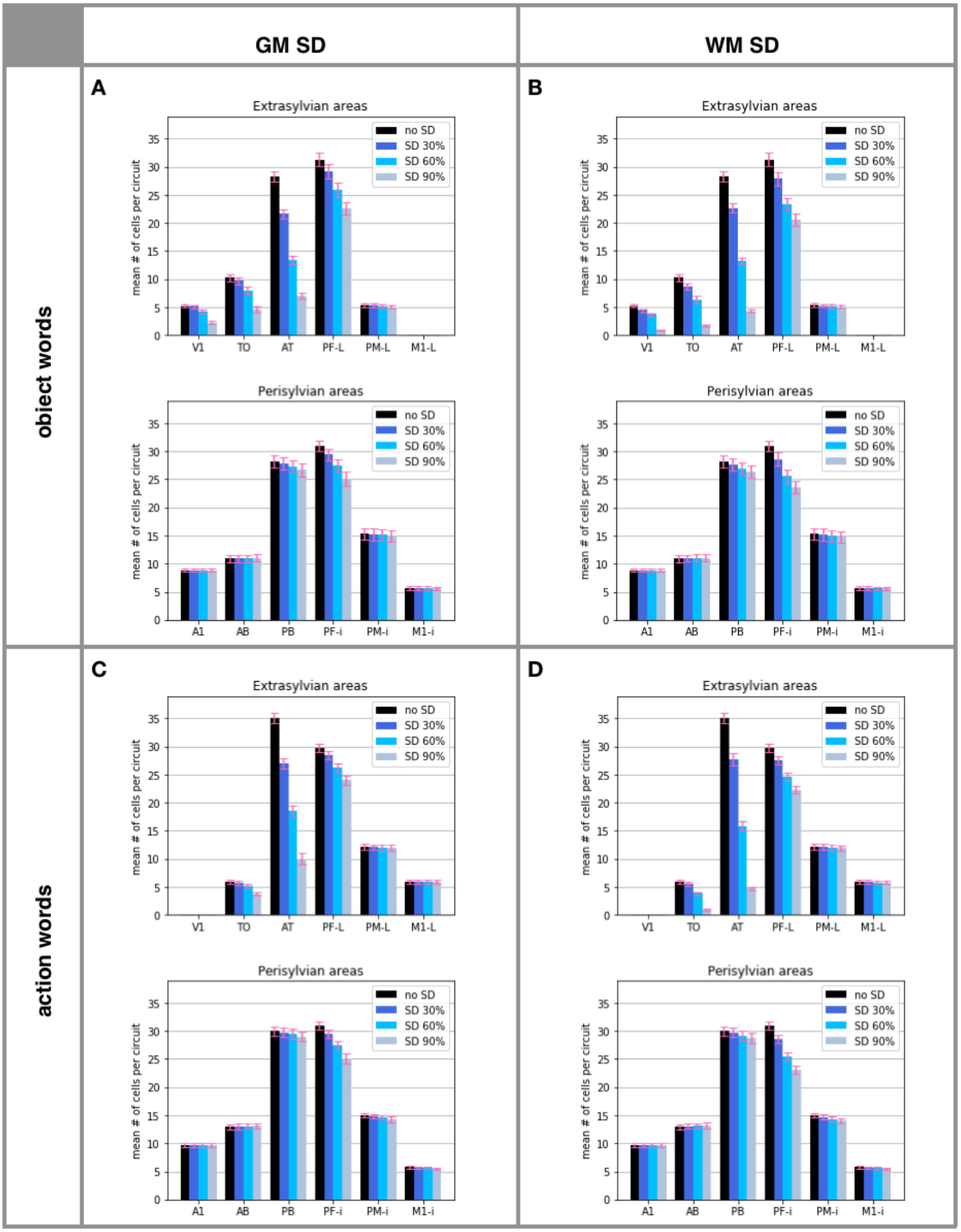

Figure 10. Activation of cell circuits after the recognition procedure during SD progression in four conditions: two word types (object- and action-related words) and two SD types (GM SD and WM SD); for extrasylvian (upper) and perisylvian (lower) areas; averaged across 13 nets. On the ordinate axis: the mean number of cells per activated circuit. Error bars indicate standard error of the mean.

interaction of the factors SDType and Severity ($F_{3,36} = 36.23$, $p < 0.001$). These results again confirm that there are differences between extrasylvian and perisylvian areas.

To explore these differences further, we calculate the total number of cells in activated circuits in all six extrasylvian areas together and the total number of cells in all six perisylvian areas together and separate four conditions: object versus action and GM SD versus WM SD. We check how the number of cells in circuits after recognition procedure differs with SD progression — from no SD through 30% and 60% SD to 90% SD — in four conditions separately (see Figure 11). Note that here we report ratio which is calculated as the number of cells in the activated circuit in the SD case divided by the number of cells in the activated circuit in no SD case. From the visual inspection of Figure 11, we can note that in all four conditions activated circuits in extrasylvian areas decline dramatically with SD severity increase, while activated circuits in perisylvian areas are more preserved.

We perform 2-way RM ANOVAs for each of the four conditions separately with factors ExtraPeri and Severity, to confirm our observations from the visual inspection. In all four ANOVAs interaction is significant (for four conditions: $F_{3,36} > 307.37$, $p < 0.001$). Then, we perform dependent samples t-tests to ensure that number of cells in circuits after recognition procedure differs significantly between no SD and 90% SD cases for each of eight conditions separately (extra/peri x object/ action x GM/WM). We also perform *post hoc* dependent samples t-tests to ensure that the number of cells in activated circuits in extrasylvian areas significantly differs from the number of cells in activated circuits in perisylvian areas in the most severe SD case (90% SD) for each of four conditions separately (object/action x GM/WM). We make Bonferroni correction for all 12 comparisons (p-value threshold is 0.0042). In the first eight comparisons, we find more cells in activated circuits for no SD case than for 90% SD case (for all eight conditions: $t_{12} > 10.71$, $p < 0.001$). In another four comparisons, we find that activated circuits in extrasylvian areas are significantly more damaged than activated circuits in perisylvian areas (for all four conditions: $t_{12} > 17.88$, $p < 0.001$).

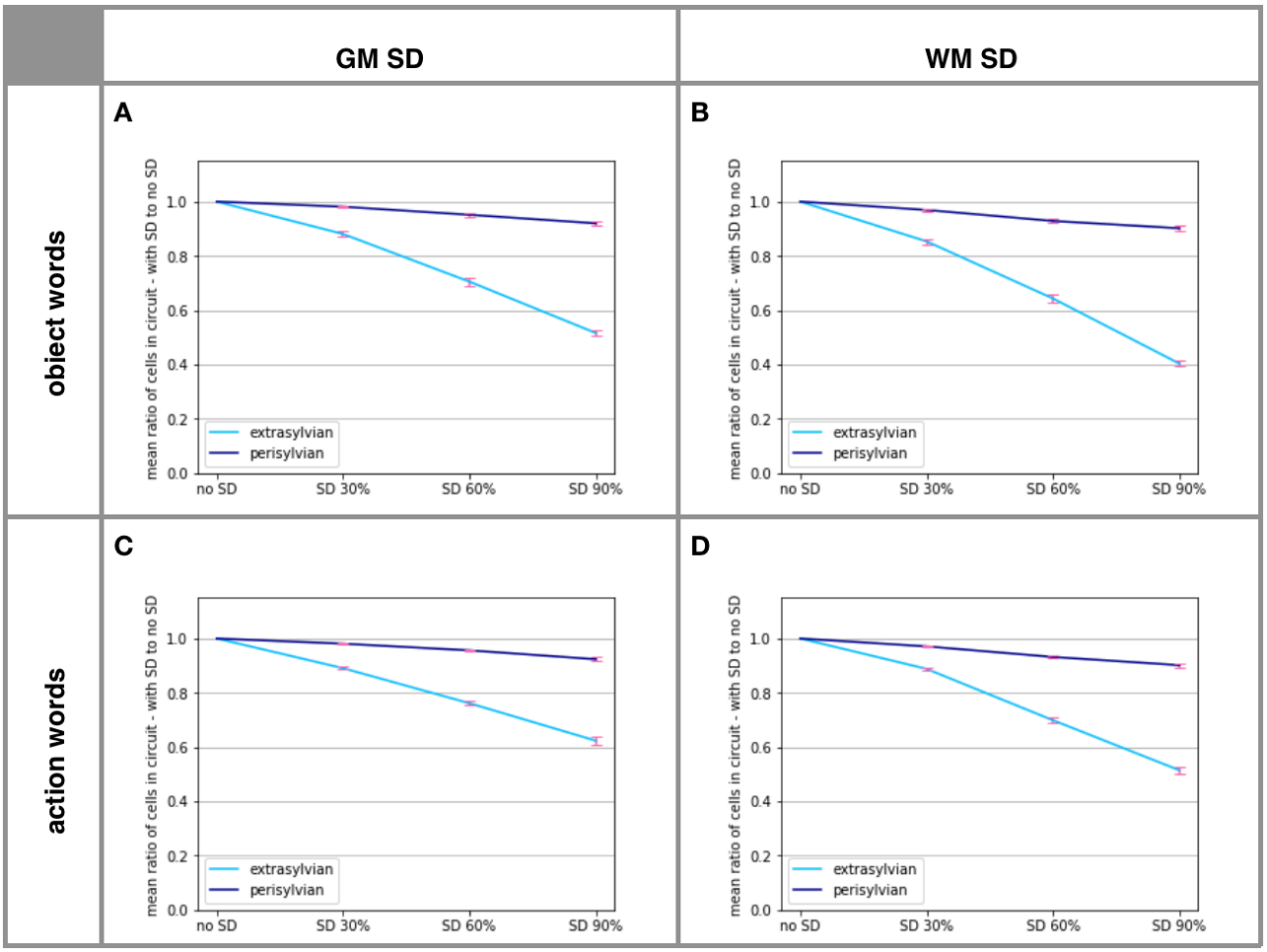

Figure 11. The mean ratio of cells in the activated (by recognition procedure) circuit during SD progression in four conditions: two word types (object- and action-related words) and two SD types (GM SD and WM SD); for all extrasylvian areas together (light blue) and all perisylvian areas together (dark blue); averaged across 13 nets. The ratio is calculated as the total number of cells in the activated circuit for six areas together in the SD case divided by the total number of cells in the activated circuit for six areas together in no SD case. Error bars indicate standard error of the mean.

Combining this graphical and statistical evidence together, we infer that in all four conditions (object/action x GM/WM) the total number of cells in activated (by recognition procedure) circuits in extrasylvian areas declines dramatically during SD severity increase: the drop is 48.4% between 90% SD case and no SD case (this number is averaged across four conditions). While in perisylvian areas the total number of cells in activated (by recognition procedure) circuits also declines, but in a

much more preserved way: the drop is 8.7% between 90% SD case and no SD case (this number is also averaged across four conditions). These findings confirm our main hypothesis: activation of cell circuits after recognition procedure decrease dramatically in extrasylvian areas with SD progression (which reflects the decile in recognition abilities of SD patients), while in perisylvian areas they are much less damaged (which reflects better-preserved repetition abilities of SD patients).

It should be noted that even within extrasylvian or perisylvian zones SD damage varies between different areas. We perform *post hoc* dependent samples t-tests to check whether the number of cells in circuits for each area differs between no SD and 90% SD cases for each of four conditions separately — object versus action and GM SD versus WM SD. We make Bonferroni correction for 48 comparisons. Differences are presented in Figure 12 (statistically significant differences are marked with asterisks). In extrasylvian areas, SD leads to a strong decline in V1, TO, AT and PF_L areas, while PM_L and $M1_L$ remain intact. In perisylvian areas, SD leads to a strong decline in PF_i and to a mild but statistically significant decline in PB and PM_i, while other areas are intact.

4.3 Semantic Dementia: Exploratory Findings for Further Experiments

While investigating the main hypothesis, we noticed that object-related concepts decline more dramatically than action-related concepts with SD progression (see Figure 13).

To test this observation statistically, we calculate the total number of cells in activated (by recognition procedure) circuits in all six extrasylvian areas together and run two 2-way RM ANOVAs (for GM SD and WM SD separately) with factors WordType and Severity. We choose only extrasylvian areas for this analysis as from the results above we can infer that these areas reflect the dramatic decline in concepts' recognition abilities with SD progression. ANOVA for GM SD shows significant interaction of two factors ($F_{3,36} = 7.04$, $p = 0.0012$). For GM SD, there is 48.3% decline of object-related words recognition in 90% SD condition, compared to no SD, which is significantly different than 37.4% decline of action-related words recognition ($t_{12} = 4.61$, $p = 0.001$). ANOVA for WM SD only shows a trend toward

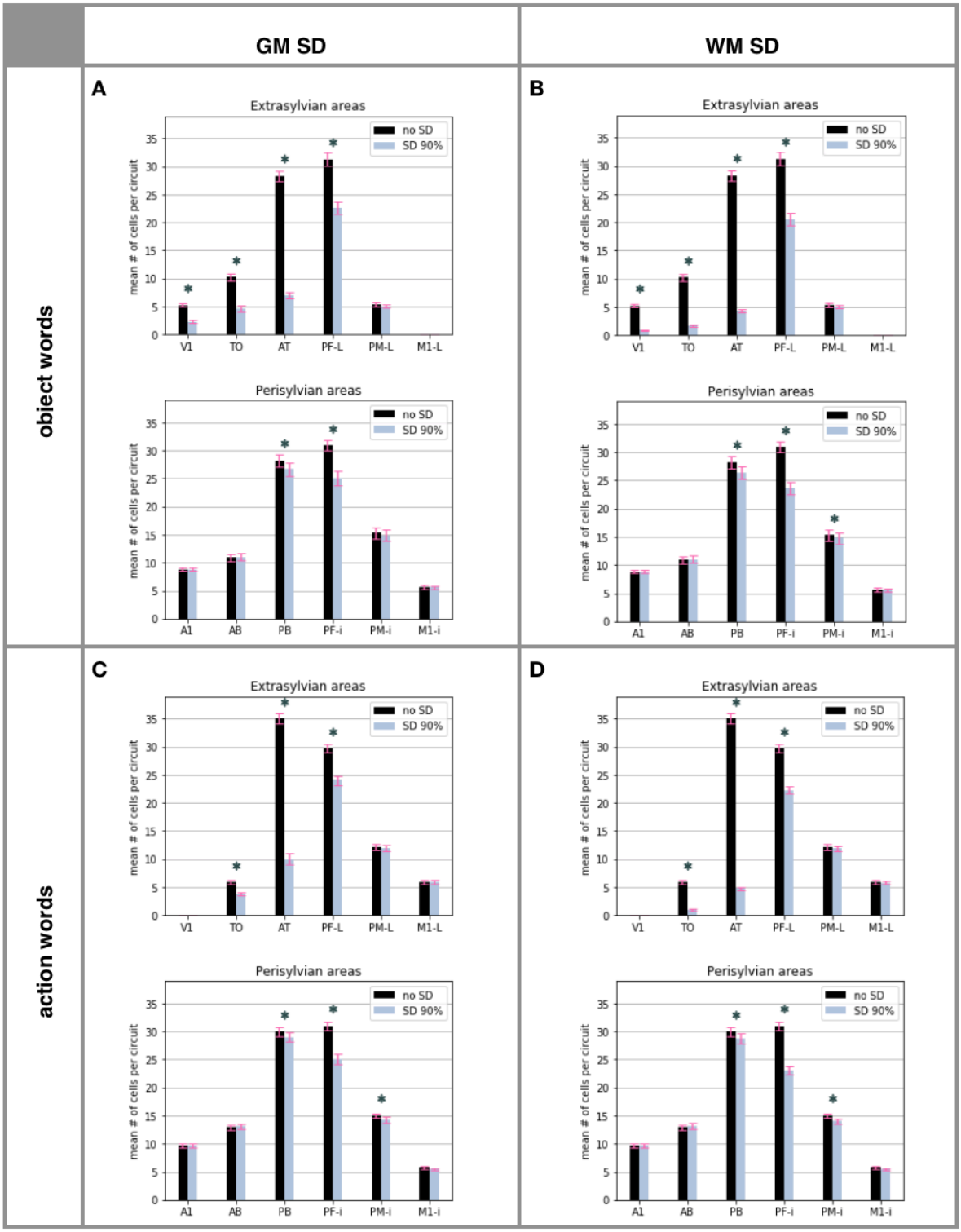

Figure 12. Activation of the cell circuits after the recognition procedure in two cases — no SD (black) and 90% SD (gray) — in four conditions: two word types (object- and action-related words) and two SD types (GM SD and WM SD); for extrasylvian (upper) and perisylvian (lower) areas; averaged across 13 nets. On the ordinate axis: the mean number of cells per activated circuit. Error bars indicate standard error of the mean.

significance ($F_{3,36} = 3.04$, $p = 0.067$). Nevertheless, it shows significance of both main factors WordType ($F_{1,12} = 18.28$, $p = 0.001$) and Severity ($F_{3,36} = 1578.12$, $p < 0.001$). Furthermore, for WM SD, we can also infer that there is 59.5% decline of object-related words recognition in 90% SD condition, compared to no SD, which is significantly different than 48.5% decline of action-related words recognition ($t_{12} = 7.82$, $p < 0.001$).

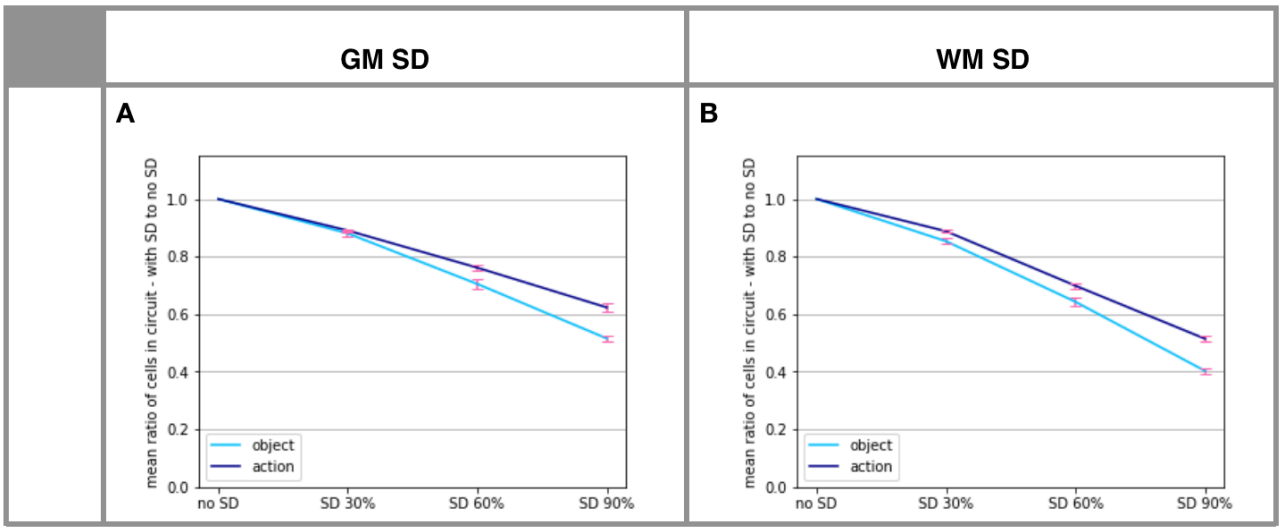

Figure 13. Comparison of the mean ratio of cells in activated (by recognition procedure) circuit during SD progression. Comparison is performed between two word types — object-related words (light blue) and action-related words (dark blue) — in two conditions: two SD types (GM SD and WM SD); for extrasylvian areas only (the total number of cells is calculated for six areas); averaged across 13 nets. The ratio is calculated as the total number of cells in the activated circuit for six areas together in the SD case divided by the total number of cells in the activated circuit for six areas together in no SD case. Error bars indicate standard error of the mean.

While investigating the main hypothesis, we also noticed that recognition abilities decline more dramatically in WM SD case than with GM SD case (see Figure 14).

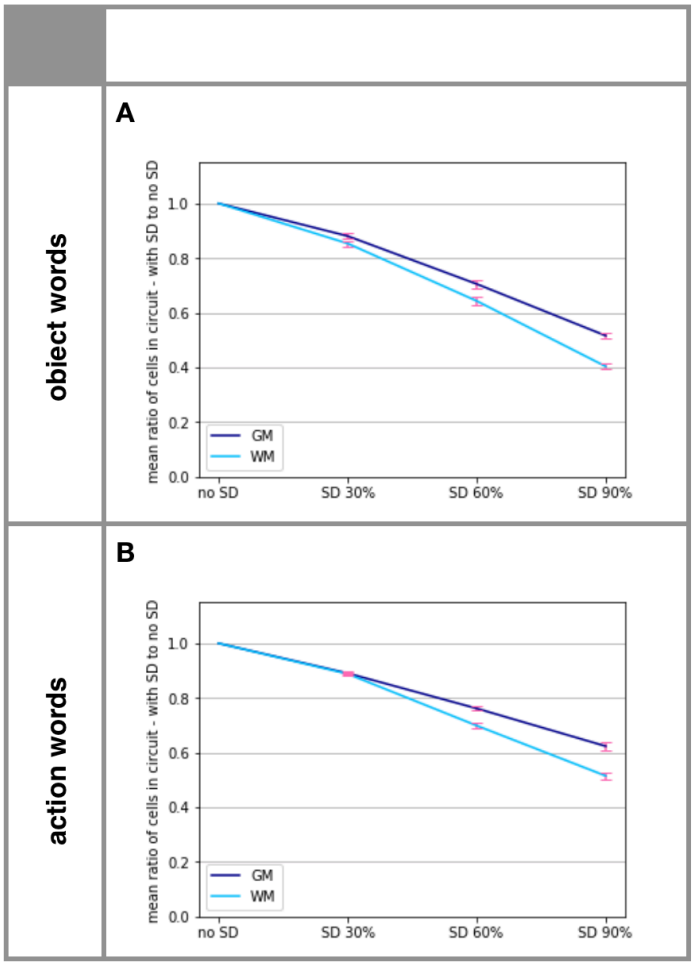

Figure 14. Comparison of the mean ratio of cells in activated (by recognition procedure) circuit during SD progression. Comparison is performed between two SD types — WM SD (light blue) and GM SD (dark blue) — in two conditions: two word types (object- and action-related words); for extrasylvian areas only (the total number of cells is calculated for six areas); averaged across 13 nets. The ratio is calculated as the total number of cells in the activated circuit for six areas together in the SD case divided by the total number of cells in the activated circuit for six areas together in no SD case. Error bars indicate standard error of the mean.

To test this observation statistically, we calculate the total number of cells in activated (by recognition procedure) circuits in all six extrasylvian areas together and run two 2-way RM ANOVAs (for object-related and action-related words separately) with factors SDType and Severity. ANOVA for object-related words shows significant interaction of two factors ($F_{3,36} = 28.52$, $p < 0.001$). For object-related

words, there is 59.5% decline of recognition abilities in 90% WM SD condition, compared to no SD, which is significantly different than 48.3% decline of recognition abilities in 90% GM SD condition. ANOVA for action-related words also shows significant interaction of two factors ($F_{3,36} = 33.04$, $p < 0.001$). For action-related words, there is 48.5% decline of recognition abilities in 90% WM SD condition, compared to no SD, which is significantly different than 37.4% decline of recognition abilities in 90% GM SD condition.

Both results are exploratory, as we did not have any predefined hypotheses about comparing object-related and action-related word degradation or WM SD and GM SD. Moreover, in the existing literature there are no experiments with SD patients that check precisely this and that can confirm or contradict our results. Thus, this provides the room for the further experiments and for developing of the model-experiment loop.

Chapter 5. Discussion

5.1 Semantic Dementia Model

In this study, we implement the semantic dementia mechanism — degradation of the gray and white matter of ATL — into the baseline model which is one of the most biologically precise neuronal models of the semantic system that exist to date (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018). We believe that creating the SD model is beneficial for better understanding of the semantic system organization, as SD is a key semantic system disease.

Using our model, we confirm the initial hypothesis that activation of concept circuits in extrasylvian (semantic) areas is dramatically disturbed with SD progression, while activation of concept circuits in perisylvian (language) areas changes very slightly. We believe that these neurological changes can explain the dynamics of SD patients — they have strong decline of recognition abilities, however, repetition abilities remain better preserved. This dissociation between perisylvian and extrasylvian areas can be checked further with neuroimaging methods applied to SD patients. Therefore, we can create a model-experiment loop by getting more precise predictions from the model to compare them against experimental data and then improve the model using this data.

Furthermore, we get predictions regarding the question of whether SD patients have category-specific or category-general recognition problems. Some researchers suggest that SD leads to category-general deficits (Lambon Ralph et al., 2007; Patterson, 2007; Machery, 2016); however, in our model, we show that recognition of object-related words decreases more sharply than recognition of action-related words with SD progression. We suggest that this effect appears as ATL is located closer to and has stronger connections with the visual areas than the motor areas. Although there were no experiments that check precisely this dynamics before, there are some clues in the existing literature that words strongly-dependent on visual features should be damaged more dramatically than others (see our review in section *2.3 Semantic Dementia*). Therefore, our model gives precise predictions that can be

checked in further experiments. This again can be used to create a better model-experiment loop.

Finally, we use two types of matter degradation in this model: white matter and gray matter degradation. Although today there are no experiments with SD patients that compare these two degradation types, from our model we can at least make a theoretical prediction about this. Our results show that white matter degradation leads to a more severe recognition decline than the same loss of the gray matter. In further research, our model can potentially be used as a tool to create the net which exactly replicates a particular patient's white and grey matter degradation severity. Then, results from patients and nets can be compared, which again creates a better model-experiment loop. Alternatively, results from the matched nets can be used to make better predictions about the future progression of the disease in the particular patients.

5.2 Alternative Explanation of Data Used in Arguments against Hybrid Theories

Another goal of this paper is to discuss different theories of semantic system organization and to show that hybrid theories suggest the most promising approach (see our review in section *2.1 Theories of Semantic System Organization*). Therefore, we build our model within the hybrid paradigm: primary perceptual areas are particularly important to provide sensory-motor co-experience during concept learning; central areas help to converge these different experiences into the united concept; emerged semantic representations are distributed between the modality-specific sensory-motor areas and the convergence hubs.

Despite the fact that there is plenty of evidence in favor of hybrid theories, there still exist arguments against it. And although all these arguments have been extensively rebuked by the authors of hybrid theories (see our review in subsection *2.1.5 Hybrid theories*), they remain to be used. Therefore, in our work we seek to further these counterarguments in favor of hybrid theories and to illustrate them more rigorously by the means of modeling.

In the first such argument, ATL, degradation of which leads to semantic system deterioration during SD, is considered as the place where concepts should be stored

in the amodal way (Patterson et al., 2007; Machery, 2016). However, as we show in our SD model, concept circuits are distributed across different areas — from primary sensory-motor areas to convergence hubs (see Figure 10) — and all of them are important for concept representations. Therefore, our results indicate that ATL is not the storage facility for amodal concept representations, but only one part of the storage facility, the convergence hub which connects different modalities, and the path through which activation spreads from the language areas to the semantic areas during word recognition. And when ATL is damaged, recognition abilities decline, due to all these factors.

In the second argument, it is suggested that, if there exist patients with lesions in the sensory-motor system who have little or no problems with the semantic system, then concepts are not really grounded on the sensory-motor system (Mahon & Hickok, 2016; for review see Barsalou, 2016). To explain this argument we use our other results (currently, preliminary)[7] — four nets are initiated as described in section *3.1 Baseline Model Description* and pass through learning and adjustment procedures; 12 CAs emerge and are evaluated; then, lesions are implemented either in the primary visual area (V1) or in the primary motor area ($M1_L$) by damaging 100% of cells in the area; and CAs are re-evaluated. We find significant but very mild decline of CA cells in extrasylvian areas due to lesions — only 6.7 % decline on average for the object- and action-related words (for both object- and action-related words: $t_3 > 3.32$, $p < 0.05$), while there are no changes at all in perisylvian areas due to lesions (for both object- and action-related words: $t_3 \leq 1$, $p > 0.39$). Moreover, this decline is mostly caused by the V1 and $M1_L$ CA cells disappearing, with CA cells in other areas being almost fully preserved (see Figure 15). Therefore, our preliminary results show that lesions in primary sensory-motor areas should not lead to the profound semantic knowledge deterioration, as CA cells are distributed between many different areas — from primary perceptual areas to convergence hubs — and CA density is higher in more central areas. Thus, to catch this deterioration we should

[7] These nets were trained with slightly different parameters than the nets where we add the SD mechanism (see parameters in Appendix); moreover, for preliminary results we train four nets only, while for the main hypothesis testing we train 13 nets.

have very sensitive tests. However, this does not mean that the sensory-motor system is not an important element of semantic representations.

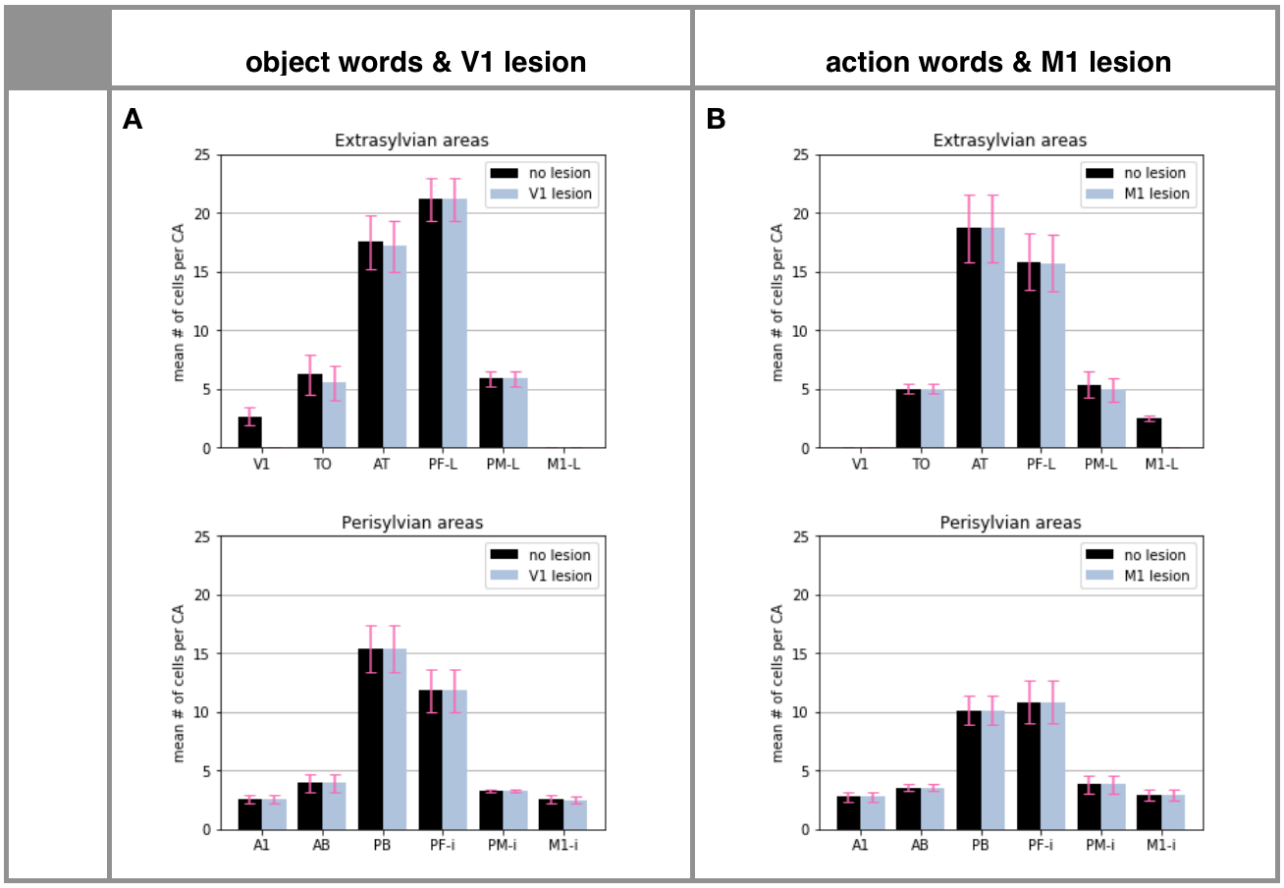

Figure 15. Preliminary results in nets with lesions in the sensory-motor system. Structure of CA in two conditions: lesions in the primary visual area V1 (left) and lesions in the primary motor area $M1_L$ (right); for extrasylvian (upper) and perisylvian (lower) areas; averaged across 4 nets. For lesions in the primary visual area (V1) only CAs for object-related words were evaluated; for lesions in the primary motor area ($M1_L$) only CAs for action-related words were evaluated. On the ordinate axis: the mean number of cells per CA. Error bars indicate standard error of the mean.

The third argument states that, if congenitally blind population does not differ from sighted population in visually-related concepts, then we do not need first-hand visual experience to create such concepts (Mahon & Hickok, 2016; Leshinskaya & Caramazza, 2016). In this study, we do not conduct any experiments with the blind

nets, however, Tomasello and colleagues (2019) did this. They used the latest iteration of our baseline model — with the improved connectivity structure and spiking neurons — and trained congenitally blind nets which do not receive any inputs to the V1 area during learning. Although they aimed to test other predictions about semantic system organization in congenitally blind population, we can infer some interesting patterns for our research from their results as well.

As we discuss in subsection *2.1.5 Hybrid theories*, to prove that conceptual knowledge is similar for congenitally blind and sighted people, it is not enough to show that the same areas are active for the same stimulus in two populations — as proponents of this argument used to do (Bedny et al., 2012; Mahon et al., 2009). We argue that the same areas can be active in both populations but the activation profiles differ. Explanation for this idea is the following: we acquire concepts through co-experience of many different sensory-motor and language inputs, therefore, inputs from the other modalities can partially compensate for the lack of the first-hand visual experience for congenitally blind individuals, thus, leading to activation of the same areas in both populations. We can see these results in the paper from Tomasello et al. (2019) (see Figure 16): in congenitally blind nets there are CA cells for object-related words in extrasylvian (semantic) areas. Although blind nets' circuits in extrasylvian areas are attenuated compared to the sighted ones, they still constitute around half of the intact circuits. Moreover, it should be noted that in this model only visual and motor modalities are used to simulate the sensory-motor system, while in reality there exist many other modalities and they are more fine-grained. For instance, the sensory-motor experience of the concept "apple" can be the following: the auditory experience of the word "apple", the visual experience of its shape and color, the gustatory experience of its taste, the olfactory experience of its smell, the motor experience of chewing it, the tactile experience of its peel, etc. Therefore, results from this paper (Tomasello et al., 2019) should be used only as a quantitative idea that concept representations can be compensated by other modalities, while more precise results should be obtained after this model is augmented with additional sensory-motor areas.

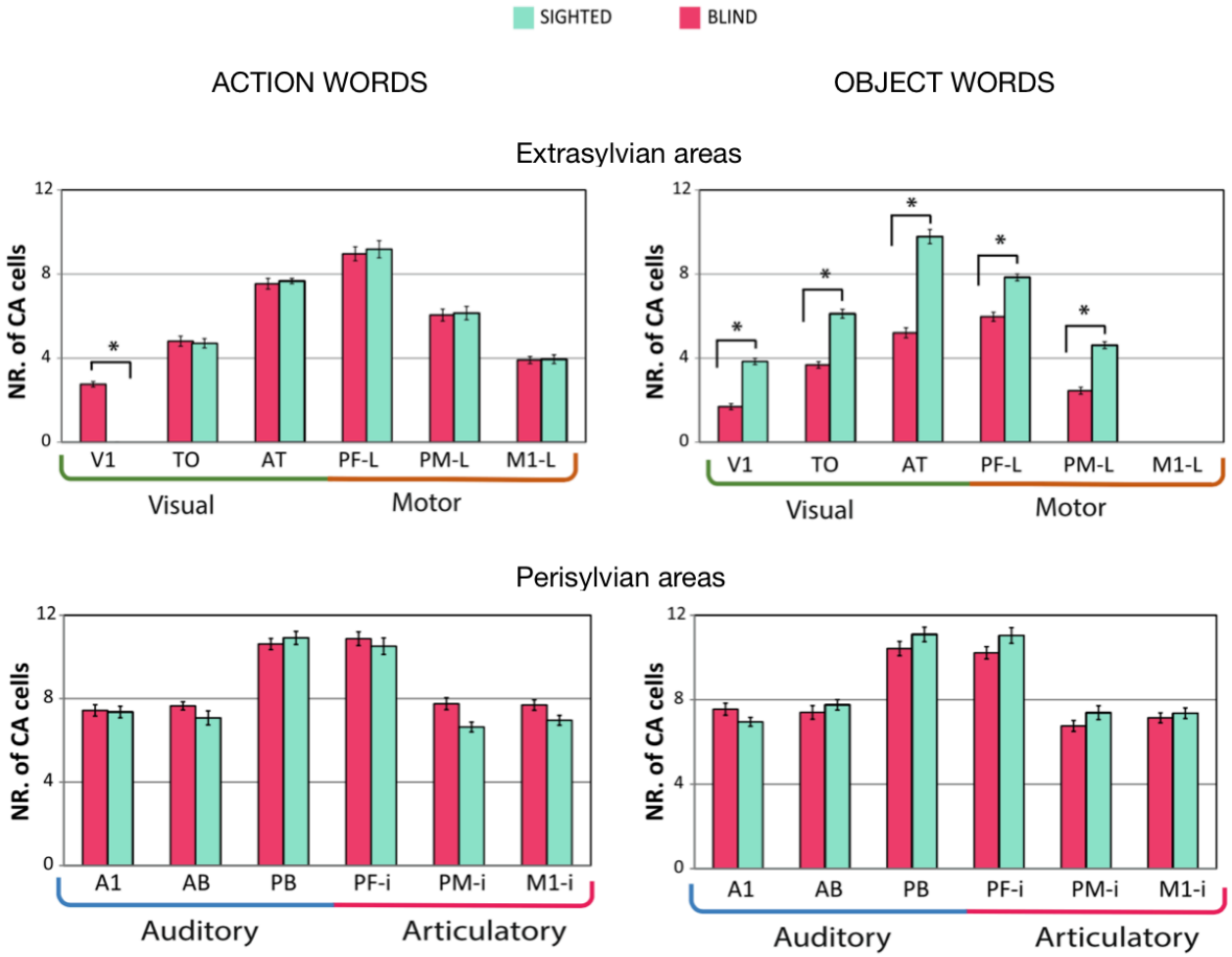

Figure 16. CA structure for the sighted (turquoise) and the blind (pink) nets in extrasylvian (upper) and perisylvian (lower) areas for two word types — action-related words (left) and object-related words (right). Statistically significant differences between sighted and blind nets within the specific area are marked with asterisks. Error bars indicate standard error of the mean. Adapted from the Visual cortex recruitment during language processing in blind individuals is explained by Hebbian learning (Figure 4), by Tomasello, R., Wennekers, T., Garagnani, M., & Pulvermüller, F., 2019, Scientific reports, 9(1), 1-16 Copyright 2019 The Author(s)

To sum up everything from the above, we believe that our model can support hybrid theories in the different controversial questions that are raised up against them.

5.3 Future of Modeling

All biologically-inspired neuronal models of semantic system that exist to date (Chen et al., 2017; Garagnani & Pulvermüller, 2016; Tomasello et al., 2018) have the same important problem: the bridge between the real data and the bio-inspired architecture of the model is currently missing. One of the plausible approaches to build this bridge is to improve and extend these biologically accurate models further. For instance, static and context-independent input patterns are usually used in such models (Chen et al., 2017; Garagnani & Pulvermüller, 2016; Tomasello et al., 2018). However, in the real world, we perceive words in dynamically changing contexts. The possibility of dynamic and context-dependent input patterns would allow to have better correspondence between the model and the real semantic system. Moreover, a very limited amount of cortex areas in the left hemisphere are usually modeled (Chen et al., 2017; Garagnani & Pulvermüller, 2016; Tomasello et al., 2018). Therefore, adding more cortex areas for both hemispheres and subcortical structures, which also contribute to the semantic memory, can be beneficial.

Nonetheless, before shifting biologically precise models to the real data domain, the important qualitative check-ups should be performed for these models in order to understand whether these best models are worthwhile to be further improved or whether there are some fatal errors in them. This is exactly what different research groups are working on now. For example, models' predictions are checked in the modeled congenitally blind individuals (Tomasello et al., 2019; Chen et al., 2017) or in the modeled patients with different lesions and diseases (Ueno et al., 2011; Chen et al., 2017).

For this reason, another goal of our study is to do such verification by adding the SD mechanism into the baseline model. We confirm that our SD model shows the specific disease dynamics: word recognition abilities decline with SD severity increase, while word repetition abilities remain better preserved, as suggested by SD patients' data (Hodges & Patterson, 2007; Montembeault et al., 2018). Therefore, we succeed in our verification — confirming that the baseline model gives qualitatively adequate results when the SD mechanism is added to it.

Last but not least, today many research groups have heated debates about different aspects of the semantic system organization (for review see Barsalou, 2016). In such situation, it may be helpful to build formal models, using which we are able to rigorously test different arguments about the dynamics of concept acquisition, concept retrieval, and other processes in the semantic system. This idea has been suggested previously by Binder (2016) and Barsalou (2016).

In this study, we present an example of such a model. We show that our model can clarify different issues regarding the semantic system organization which continuously raise controversy. Moreover, using this model we make many specific predictions that can be tested in further experiments.

Conclusion

In this study, we discuss modern theories of semantic system (Mahon & Caramazza, 2008; Mahon, 2015; Rogers et al., 2004; Patterson, Nestor, & Rogers, 2007; Gallese & Lakoff, 2005; Kiefer & Pulvermüller, 2012; Garagnani & Pulvermüller, 2016; Binder & Desai, 2011; Binder, 2016; Barsalou, 2016) and different approaches to model it (Huth et al., 2016; Anderson et al., 2016; Chen et al., 2017; Garagnani & Pulvermüller, 2016; Tomasello et al., 2019). We argue that hybrid theories — which postulate both substantial grounding in the sensory-motor system and higher-order convergence zones — are the most promising way to think about semantic representations (Kiefer & Pulvermüller, 2012; Garagnani & Pulvermüller, 2016; Binder & Desai, 2011; Binder, 2016; Barsalou, 2016). We use biologically-inspired neuronal model of the semantic system, built within this hybrid paradigm (Garagnani & Pulvermüller, 2016; Tomasello et al., 2018), as the basis, and augment it with the semantic dementia mechanism, as SD is the key disease which influences the semantic system (Hodges & Patterson, 2007; Montembeault et al., 2018; Spinelli et al., 2017).

We demonstrate that our SD model's predictions can help with a better understanding of the semantic system organization. We also show how our baseline model puts into question the most common arguments against hybrid theories. Finally, we highlight the necessity to improve such models if we want to be more productive in our research on semantic system understanding.

References

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.

Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic bulletin & review*, *23*(4), 1122-1142.

Bedny, M., Caramazza, A., Pascual-Leone, A., & Saxe, R. (2012). Typical neural representations of action verbs develop without vision. *Cerebral cortex*, *22*(2), 286-293.

Bedny, M., Koster-Hale, J., Elli, G., Yazzolino, L., & Saxe, R. (2019). There's more to "sparkle" than meets the eye: Knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition*, *189*, 105-115.

Binder, J. R. (2016). In defense of abstract conceptual representations. *Psychonomic bulletin & review*, *23*(4), 1096-1108.

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.

Bird, H., Lambon Ralph, M. A., Patterson, K., & Hodges, J. R. (2000). The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and language*, *73*(1), 17-49.

Bonner, M. F., Vesely, L., Price, C., Anderson, C., Richmond, L., Farag, C., ... & Grossman, M. (2009). Reversal of the concreteness effect in semantic dementia. *Cognitive Neuropsychology*, *26*(6), 568-579.

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.

Breedin, S. D., Saffran, E. M., & Coslett, H. B. (1994). Reversal of the concreteness effect in a patient with semantic dementia. *Cognitive neuropsychology*, *11*(6), 617-660.

Cardebat, D., Demonet, J. F., Celsis, P., & Puel, M. (1996). Living/non-living dissociation in a case of semantic dementia: a SPECT activation study. *Neuropsychologia*, *34*(12), 1175-1179.

Chen, L., Lambon Ralph, M. A., & 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.

Daniele, A., Giustolisi, L., Silveri, M. C., Colosimo, C., & Gainotti, G. (1994). Evidence for a possible neuroanatomical basis for lexical processing of nouns and verbs. *Neuropsychologia*, *32*(11), 1325-1341.

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.

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

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., Martinos, M., Woollams, A. M., Patterson, K., & Adlam, A. L. R. (2008). Repeat and point: differentiating semantic dementia from progressive non-fluent aphasia. *cortex*, *44*(9), 1265-1270.

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.

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.

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

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

Lambon Ralph, M. A., 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.

Leshinskaya, A., & Caramazza, A. (2016). For a cognitive neuroscience of concepts: Moving beyond the grounding issue. *Psychonomic bulletin & review*, *23*(4), 991-1001.

Machery, E. (2016). The amodal brain and the offloading hypothesis. *Psychonomic bulletin & review*, *23*(4), 1090-1095.

Mahon, B. Z. (2015). What is embodied about cognition?. *Language, cognition and neuroscience*, *30*(4), 420-429.

Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., & Caramazza, A. (2009). Category-specific organization in the human brain does not require visual experience. *Neuron*, *63*(3), 397-405.

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.

Mahon, B. Z., & Hickok, G. (2016). Arguments about the nature of concepts: Symbols, embodiment, and beyond. *Psychonomic bulletin & review*, *23*(4), 941-958.

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., & Lambon Ralph, M. A. (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., Cooper-Pye, E., Dine, C., Hauk, O., Nestor, P. J., & Patterson, K. (2010). The word processing deficit in semantic dementia: all categories are equal, but some categories are more equal than others. *Journal of Cognitive Neuroscience*, *22*(9), 2027-2041.

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.

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

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

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

Saysani, A., Corballis, M. C., & Corballis, P. M. (2018). Colour envisioned: Concepts of colour in the blind and sighted. *Visual Cognition*, *26*(5), 382-392.

Shepard, R. N., & Cooper, L. A. (1992). Representation of colors in the blind, color-blind, and normally sighted. *Psychological Science*, *3*(2), 97-104.

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), 1-16.

Ueno, T., Saito, S., Rogers, T. T., & Lambon Ralph, M. A. (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

Parameters of the model

$\tau_e = 2.5$	time constant for membrane potential of the e-cell **(E1.1)**
$\tau_i = 5$	time constant for membrane potential of the i-cell **(I1.1)**
$\tau_G = 12$	time constant for global inhibition **(E1.3)**
$\tau_A = 10$	time constant for averaged recent activity of the e-cell **(E2.3)**
$k_1 = 0.01$	scaling constant **(E1.1)**
$k_2 = 100\sqrt{3}$	scaling constant **(E1.1)**
$k_G = 95$	scaling constant **(E1.2)**
$\alpha = 0.01$	adaptation strength **(E2.2)**
$\eta \sim U[-0.5, 0.5]$	noise distribution **(E1.1)**
$w_{initial} \sim U[0, 0.1]$	initial weight distribution for links between two e-cells
$\Delta w = 0.0008$	change of synaptic weight during Hebbian learning
$w_{inh} \sim N(0.295, 2)$	weight distribution for links from the e-cell to a 5x5 patch of i-cells, decreasing with the distance between cells
$\theta_{pre} = 0.05$	threshold for presynaptic activity
$\theta_{post} = 0.15$	threshold for postsynaptic activity
$\gamma = 0.5$	fraction of the time-averaged output of the most active e-cell to determine CA threshold

Probability that links between two e-cells will be created follows Gaussian distribution, decreasing with the distance between two e-cells, and is equal to zero outside 19x19 patch.

$Pr_{rec} = 0.15$	central probability of Gaussian distribution for recurrent links
$Pr_{between} = 0.28$	central probability of Gaussian distribution for links between different areas
$\sigma_{rec} = 4.5$	standard deviation for Gaussian distribution for recurrent links
$\sigma_{between} = 6.5$	standard deviation for Gaussian distribution for links between different areas

---

Training parameters for SD nets from the main part of the paper.

SI = 500	sensory input amplitude
MI = 500	motor input amplitude
J = 800	links strength (between areas, recurrent and inhibitory)
NI = 25	noise intensity

Training parameters for nets with lesions from Discussion.

SI = 300	sensory input amplitude
MI = 300	motor input amplitude
J = 500	links strength (between areas, recurrent and inhibitory)
NI = 25	noise intensity

Acknowledgements

I would like to thank Alexey Guzey for editing and support.

Additional information

All tables with data, scripts for graphs generation and statistical calculations, ANOVAs tables, and t-tests results can be found here: https://github.com/ansty57/Master_Thesis_hse2020