University of Tartu
Institute of Philosophy and Semiotics

REVERSIBLE AND IRREVERSIBLE WHORFIAN EFFECTS AS EMPIRICAL EVIDENCE FOR THE SAPIR-WHORF HYPOTHESIS

Master's Thesis in Philosophy
Maksim Grigorev

Supervisor: Alexander Davies
Co-supervisor: Bruno Mölder

Tartu, 2018

Table of Contents

1. Introduction

For this MA thesis I have three goals. The first goal is to determine a clear definition of reversible Whorfian effects, which were firstly noticed by Reines and Prinz (2009). The second goal is to show that reversibility can be an item of contention, when researchers discuss whether or not empirical data can support the Sapir-Whorf hypothesis. Namely, some researchers can reject Whorfian effects as empirical evidence for the Sapir-Whorf hypothesis only because these Whorfian effects are reversible. In reality, reversibility is a new concept, which hasn't been researched well enough. That is why the point of view that reversible Whorfian effects can't support the Sapir-Whorf hypothesis should be justified. The third goal is to show that some Whorfian effects don't demonstrate properties of reversible Whorfian effects, which makes them good candidates to be identified as irreversible. All the three goals are aimed to help in philosophical research on formulating a scientifically testable thesis of the Sapir-Whorf hypothesis. The first goal separately also will help to identify reversible Whorfian effects more easily in future research on the Sapir-Whorf hypothesis in case, if such research will include usage of empirical data.

Roughly, *the Sapir-Whorf hypothesis* is the thesis that *language somehow affects thought,* and this thesis has a long research tradition. Nowadays, a lot of experiments in the Sapir-Whorf hypothesis provide empirical data, but researchers have no common opinion about whether or not such empirical data can support the Sapir-Whorf hypothesis. Authors simply don't design their experiments to prove or disprove the Sapir-Whorf hypothesis. Nonetheless, researchers still take these experiments to be related to the Sapir-Whorf hypothesis. This happens because experiments in the Sapir-Whorf hypothesis explore more specific relations between some properties of language (e.g. grammatical gender, system of spatial orientation, color categorization etc.) and some properties of thought (e.g. cognitive abilities, like memorizing, color discrimination, choice making etc.):

> […] the Sapir-Whorf hypothesis turned into series of exact and rigorous scientific experiments. […] modern linguists […] provide absolutely concrete, narrow experiments, comparing the reaction of representatives of different peoples, who speak different languages, and connect results with their languages, grammar arrangement and semantics arrangement. (Krongaus 2012, 10:00 – 11:15)

These relations between some properties of language and some properties of thought are obviously related to the Sapir-Whorf hypothesis, but the claim that language affects thought is much broader than e.g. relations between grammatical gender and memorizing. In other words, explored relations between some properties of language and some

properties of thought can be not enough to prove the thesis that language affects thought. Therefore, the thesis of the Sapir-Whorf hypothesis should be obviously clarified to deal with empirical data. Some scientists suppose that *a scientifically testable thesis* should be firstly formulated to test whether or not some existing or following empirical data can support the Sapir-Whorf hypothesis (van Troyer 1994). I'm pretty sure that research of *reversibility* is necessary for formulating of the scientifically testable thesis.

The problem with the claim *language affects thought* is that every word in this claim is ambiguous and it's not explicitly clear what words "language", "affects" and "thought" mean. Therefore, researchers of the Sapir-Whorf hypothesis tried to clarify its thesis long before the necessity to formulate the scientifically testable thesis appeared. As a result, different interpretations of the Sapir-Whorf hypothesis appeared. The two most common interpretations of the Sapir-Whorf hypothesis are called *the strong form* and *the weak form*. The strong form of the Sapir-Whorf hypothesis states that *language determines* or *shapes thought*, the weak form means that *language influences thought* (Krongauz 2012, 00:00-02:00), (Ahearn 2011, 69). These two forms of the Sapir-Whorf hypothesis basically differ in the point of view on whether or not *language is necessary for thought*, i.e. the strong form of the Sapir-Whorf hypothesis means that language is necessary for thought, while the weak form means that language isn't necessary for thought.[1] Unfortunately, neither the strong nor the weak form of the Sapir-Whorf hypothesis can help a lot in formulating of the scientifically testable thesis because both interpretations are still ambiguous. Nonetheless, nowadays the strong form is mostly rejected by researchers of the Sapir-Whorf hypothesis (Ahearn 2011, 69), while the weak form is still being discussed by researchers. This means that if researchers try to find out whether or not empirical data supports the Sapir-Whorf hypothesis, then they test exactly its weak form. But even the claim of the weak form is ambiguous and should be clarified to be tested by empirical data.

This problem is obviously caused by the absence of a scientifically testable thesis. Van Troyer rightly noticed that: "[Sapir-Whorf hypothesis] does not exist as a scientifically testable thesis. […] the Sapir-Whorf "Hypothesis" exists only as a notion" (van Troyer 1994, 163). This means that researchers have no unified interpretation of the Sapir-Whorf hypothesis to test. Nevertheless, some researchers suppose that "[…] this philosophical argument around the relation of thoughts and language(s) shall be finally

1. I will justify this claim in chapter three, where I will show that two groups of authors, who tried to find out whether or not the Sapir-Whorf hypothesis can be supported by empirical data, rejected the strong form of the Sapir-Whorf hypothesis exactly because of the claim that language is necessary for thought.

settled through empirical and experimental evidence rather than analytic considerations" (Pöhls 2013, 99). In that case, researchers should find out whether or not empirically explored relations between some properties of language and some properties of thought can be interpreted as the claim that language affects thought. Without the unified scientifically testable thesis, researchers test different interpretations of the Sapir-Whorf hypothesis and get different answers to the question whether or not empirical data can support the Sapir-Whorf hypothesis. Van Troyer commented on this situation in the following way: "As a result, all studies which have attempted to interpret empirical data according to the hypothesis are either flawed or invalid because they have tested something other than the hypothesis" (van Troyer 1994, 163). It is precisely philosophers who seem to be able to solve the problem of the scientifically testable thesis, because this obviously needs to involve theoretical investigations and philosophers are always good in developing new theories.

There is even an attempt to formulate a guideline for empirical research of the Sapir-Whorf hypothesis by a philosopher, who wrote a paper "Testing the untestable? Guidelines for advancing empirical research in the area of Linguistic Relativity" (Pöhls 2013). In this paper, Pöhls formulated the list of demands for empirical research, which should help to develop strong evidence for the Sapir-Whorf hypothesis. Nevertheless, the scientifically testable thesis seems to be far away from its final formulation. That is why researchers still try to test different interpretations of the Sapir-Whorf hypothesis by empirical data. My MA thesis should help in future formulation of the scientifically testable thesis by research of *reversibility* and its properties. Also, I will present a couple of Whorfian effects, which are candidates to be called *irreversible*. Future research of irreversible Whorfian effects is also important for formulation of the scientifically testable thesis.

First of all, I will explain what reversibility and its properties are in the second chapter. Roughly, *reversibility* is a property of *Whorfian effects* to be *reversible*, i.e. to be erased under some circumstances. "Whorfian effects" is the term used by Reines and Prinz (2009) to name these effects of some language properties on some cognitive abilities, which I already mentioned. Reversibility and circumstances, under which Whorfian effects become reversible, aren't well researched. Reines and Prinz just noticed in their paper that "Whorfian effects are often reversible" (ibid. 1028) and introduced two experiments: the first experiment explored the Whorfian effect in spatial orientation and the second experiment reversed this Whorfian effect under slightly changed circumstances. Reines

and Prinz didn't provide a definition of reversible Whorfian effects, didn't make a list of reversible Whorfian effects and didn't notice whether some other Whorfian effects exist. The thing is that reversible Whorfian effects weren't the topic of the paper by Reines and Prinz, who actually stressed that even reversible Whorfian effects can support interpretations of the Sapir-Whorf hypothesis, which were proposed in their paper (*Habitual* and *Ontological Whorfianism*). This means that reversibility isn't a problem for Reines and Prinz in their attempt to support the Sapir-Whorf hypothesis by empirical data. Another group of researchers represented by Weiskopf and Adams (2015) conversely rejected all the reversible Whorfian effects as evidence for the Sapir-Whorf hypothesis, but reversibility also wasn't the topic of their research and they didn't provide any definition. My goal for the second chapter is to recap the most important reversible Whorfian effects, to show how they were reversed and to define properties of these reversible Whorfian effects. Defining of reversible Whorfian effects' properties will help me in the third chapter to answer the question whether or not these properties can be avoided, i.e. whether or not *irreversible* Whorfian effects exist. Moreover, defined properties of reversible Whorfian effects will help in future tests of empirical data. Nowadays, researchers try to analyze every experiment in the Sapir-Whorf separately, without any methodology, as was done by Weiskopf and Adams. Again and again, Weiskopf and Adams had to prove that the next experiment should be also rejected. Their analysis of the experiments was correct, but they actually rejected all the presented Whorfian effects for the same reason, i.e. because they were reversible. Defined properties of reversible Whorfian effects will at least save time and researchers will identify much more simply whether or not some Whorfian effects are reversible. In that case researchers of the Sapir-Whorf hypothesis will just have to find out whether or not reversible Whorfian effects can support the Sapir-Whorf hypothesis.

In the third chapter of my MA thesis, I will show that research of reversibility should help to avoid misunderstanding in formulating of the scientifically testable thesis. For now, different understanding of reversibility can be a reason why some researchers reject Whorfian effects as empirical evidence of the Sapir-Whorf hypothesis, while other researchers accept them. For example, Reines & Prinz and Weiskopf & Adams sometimes even used the same experiments in their arguments, but still had different points of view on whether or not explored in these experiments Whorfian effects can support the Sapir-Whorf hypothesis. Both groups of authors also used really similar interpretations of the Sapir-Whorf hypothesis to test whether or not these interpretations can be supported by

empirical data. In other words, Reines & Prinz and Weiskopf & Adams used the same Whorfian effects and interpretation of the Sapir-Whorf hypothesis for tests, but Reines and Prinz accepted these Whorfian effects, while Weiskopf and Adams rejected them as empirical evidence to support the Sapir-Whorf hypothesis. Therefore, my goal for the third chapter is to show that different understandings of reversibility can be the reason for different points of view on whether or not empirical data can support the Sapir-Whorf hypothesis. To realize this goal, I should firstly clarify the terminology, which was used by the authors. The problem is that Reines & Prinz and Weiskopf & Adams used totally different terminology to describe their interpretations of the Sapir-Whorf hypothesis and reversibility. Moreover, the concept of reversible Whorfian effects is totally new and there is no common definition of it. Reines and Prinz briefly described it, but didn't formulate it. Weiskopf and Adams were proving that all the Whorfian effects used in their paper are reversible, but didn't notice that this is a systematic problem and didn't name it in any way. Therefore, somebody could think that Reines & Prinz and Weiskopf & Adams were talking about different things in their papers. That is why I will firstly clarify some basic terminology in subsection 3.1 to be sure that Reines & Prinz and Weiskopf & Adams used for tests the same interpretation of the Sapir-Whorf hypothesis and that they all discussed exactly reversible Whorfian effects (not some other effects). Otherwise, I would not have any reason to compare works by Reines & Prinz and Weiskopf & Adams, if they were talking about totally different things. Then I will justify in subsection 3.2 that Reines and Prinz agreed that reversible Whorfian effects are evidence for the Sapir-Whorf hypothesis. In subsection 3.3 I will show that Weiskopf and Adams rejected reversible Whorfian effects as evidence for the Sapir-Whorf hypothesis. One more topic for subsection 3.3 is to show that Weiskopf and Adams actually proved that the Whorfian effects discussed by them are reversible, but they didn't prove that reversible Whorfian effects should be rejected as empirical evidence for the Sapir-Whorf hypothesis. As a result, the third chapter will show that reversibility should be researched to avoid misinterpretation of Whorfian effects in formulating of the scientifically testable thesis.

In the fourth chapter, I will stress that some Whorfian effects can avoid the features of reversible Whorfian effects, which I will define in the second chapter. My point is that all the discussed by Reines & Prinz and Weiskopf & Adams Whorfian effects are reversible and don't need to be additionally interpreted separately. Instead of this, future research should find out whether or not reversible Whorfian effects can support the Sapir-Whorf hypothesis as a special type of empirical data. But some researchers seem already to

reject reversible Whorfian effects as evidence for the Sapir-Whorf hypothesis, as was done by Weiskopf and Adams. Therefore, it's also crucial for formulating of the scientifically testable thesis of the Sapir-Whorf hypothesis to answer the question whether or not irreversible Whorfian effects exist. In the fourth chapter, I will represent a set of experiments, which probably explored *irreversible Whorfian effect*. This set of experiments was presented by Deutscher (2010) as argument for the Sapir-Whorf hypothesis, but wasn't justified enough. As for Reines & Prinz or Weiskopf & Adams, they didn't discuss this set of experiments for some reason and this is confusing. Nonetheless, Reines and Prinz mentioned the experiment by Winawer et al. (2007) from the set, but didn't specify whether or not this experiment demonstrated an irreversible Whorfian effect. My job for the fourth chapter is to find out whether or not this Whorfian effect is really irreversible. The fourth chapter should complete my research on reversibility and this will be obviously useful for formulating of the scientifically testable thesis, when I also answer whether or not some other Whorfian effects exist.

In the conclusion, I will finally postulate what reversibility is according to its features. I will also summarize some features of my candidates to be identified as irreversible Whorfian effects. I guess that even if my candidates for irreversible Whorfian effects can be somehow reversed, nevertheless, they can become much stronger proof of the Sapir-Whorf hypothesis for researchers, who rejected reversible Whorfian effects as empirical evidence for the Sapir-Whorf hypothesis.

2. Reversible Whorfian effects and their properties

In this chapter, my main goal is to recap the most important experiments in the Sapir-Whorf hypothesis, to show how Whorfian effects in these experiments were or can be reversed and to define properties, which made these Whorfian effects reversible. Defined properties of reversible Whorfian effects will help in future research to detect reversible Whorfian effects much more simply. Moreover, these properties will help me to split obviously reversible Whorfian effects from candidates to be identified as irreversible Whorfian effects. All of this should also help in formulating a scientifically testable thesis of the Sapir-Whorf hypothesis.

I wrote that I should recap the most important experiments, because the list of the experiments in the Sapir-Whorf hypothesis is actually much longer. Reines & Prinz and Weiskopf & Adams, whose works are crucial for my MA thesis because only these authors somehow described reversible Whorfian effects for today, marked results of many experiments as ambiguous. Therefore, I will recap only those experiments, which don't raise any doubts in their results. Reines & Prinz and Weiskopf & Adams often discussed the same Whorfian effects as real candidates for those which might support the Sapir-Whorf hypothesis. Moreover, Reines & Prinz and Weiskopf & Adams presented two different points of view on whether reversible Whorfian effects can support the Sapir-Whorf hypothesis. This is why I based my list of the most important experiments on the opinion of Reines & Prinz and Weiskopf & Adams.

I already mentioned in the introduction what the terms "Whorfian effects" and "reversibility" roughly mean. In subsection 3.1 I will recap all the terminology and explanations, which were presented by Reines & Prinz and Weiskopf & Adams. My aim in recapping their terminology is to verify whether they really discussed the same interpretation of the Sapir-Whorf hypothesis and reversibility, while they used totally different terminology and sometimes very ambiguous explanations. In reality, neither Reines & Prinz nor Weiskopf & Adams give any explicit definition of "Whorfian effects" and "reversibility". Therefore their terminology will not help me to somehow define properties of reversible Whorfian effects. All my reader needs to know for now is that to show whether or not some Whorfian effect is really reversible, I should show how this Whorfian effect was explored and how it was reversed, i.e. erased.

Of course, it's also a question of how researchers can be sure that a given Whorfian effect was really reversed. From the papers by Reines & Prinz and Weiskopf & Adams, it's

intuitively clear that Whorfian effects are reversible, if they disappear in a slightly modified situation. For example, some Whorfian effects are explored in some experiment, but disappear when some properties of the experiment are changed slightly. And there is also a question of how it can be really the same Whorfian effect, when some properties of the experiment were changed slightly. Nevertheless, Reines & Prinz and Weiskopf & Adams introduced their arguments to explain why they defined in their papers their presented Whorfian effects as reversible. Therefore, in this chapter I will firstly recap all the necessary experiments and the way that they were or can be reversed. Only then will I define properties of reversible Whorfian effects, including my answers to the questions on how I can verify that in every case it was the same Whorfian effect and it was really reversed.

In accordance with my goals for this chapter, I will recap reversible Whorfian effects in subsections 2.1-2.3 and then I will define properties of reversible Whorfian effects in subsection 2.4. As I already mentioned in the introduction, Reines & Prinz introduced in their paper only one example of a reversible Whorfian effect, while Weiskopf & Adams provided a long list of reversible Whorfian effects. Therefore I will also use the structure of Weiskopf & Adams to divide Whorfian effects into three groups: ontology, space and gender. The only difference in my structure will be that I will present groups of Whorfian effects in different sequences: space in subsection 2.1., ontology in subsection 2.2 and gender in subsection 2.3. I decided to swap space and ontology in my MA thesis because, in my opinion, Whorfian effects in space are reversed in the most obvious way (empirically, by a counter experiment) and this would be the best example to start from. In the end of this chapter, I will summarize all the results, which I will find by researching reversible Whorfian effects.

2.1. Reversible Whorfian effects in spatial orientation

Weiskopf and Adams used the same strategy to reject all the three arguments: space, ontology and gender. Their strategy was to show that Whorfian effects appear only in too narrow cases and this is too weak evidence to claim that language affects thought. That is, these Whorfian effects are reversible as Reines and Prinz would say. If reversible Whorfian effects can really appear only in some narrow experimental situation, then their influence is too weak to take this into account in the dispute about whether or not empirical data can support the Sapir-Whorf hypothesis. Therefore, Weiskopf and Adams try to imagine some situation where people should demonstrate some Whorfian effect, but for some reason

failed. Whorfian effects in spatial orientation seem to be the best example, because their reversibility was proven empirically by counter experiment with almost the same procedure, but with a changed background.

The space argument is obviously based on the fact that some languages express spatial relations really differently, but not all of these differences can be used to detect some Whorfian effect and suggest differences in thought. In their subsection about the space argument, Weiskopf and Adams represented several experiments, but I will concentrate only on the really important experiments. For example, Weiskopf and Adams recapped some papers to discuss the fact that: "Korean ignores the English "in" and "on" as distinctions for containment, but pays close attention to whether the containment is "tight" (a cap going on a pen) or "loose" (apples in an open bowl)" (Weiskopf & Adams 2015, 258). But the problem is that Weiskopf and Adams didn't introduce any experiment to suggest that such different categorization of space can somehow influence thought. In other words, this group of papers about different spatial categories in Korean seems to be only descriptive and has nothing to detect any Whorfian effect. Reines & Prinz and Weiskopf & Adams also mentioned research in spatial orientation by Brown & Levinson (2000), but this research also seems to be only descriptive and doesn't demonstrate any Whorfian effect. Therefore, I will recap only two experiments in spatial orientation, which really explored or reversed some Whorfian effects: the experiment by Pederson et al. (2008) and the experiment by Li & Gleitman (2002). The first experiment demonstrated how different systems of spatial orientations determined participant's choice in a non-linguistic task. The second experiment showed that differences in system of spatial orientation don't influence thought constantly and people in different situations can use different systems of spatial orientation, which proves that such a Whorfian effect is reversible. Both experiments were recapped by Reines & Prinz and Weiskopf & Adams as an example of the reversible Whorfian effect. This obviously means that Reines & Prinz and Weiskopf & Adams defined reversibility basically in the same way.

In the experiments by Pederson et al. participants had to solve a simple non-linguistic task. Participants of the experiment were divided into two groups: with *absolute system* of spatial orientation and with *relative system* of spatial orientation in their mother tongue. Speakers of languages with an absolute system of spatial orientation describe the position of objects in space according to some geographical objects, e.g. by opposition uphill-downhill, by cardinal directions etc. Speakers of languages with a relative system of spatial orientation describe the position of objects in space according to themselves, e.g. by

the sides of their body: from the left side, from the right side, straight ahead etc. Even in the case, when speakers of languages with a relative system of spatial orientation seem to orient objects according to sides of some other object (building, road, river etc.), they still imagine in every case an observer according to whose position these objects are located from some side. In reality, a road doesn't have a right or left side, i.e. sides of such a road can only be actualized by a person according to a context.

In the experiment two tables were presented: the stimulus table and the recall table. Tables were parallel to each other. Firstly, every participant sat at the stimulus table. On the stimulus table there was a line of toys. Toys were chosen to have distinctive directions, e.g. it was clear which way a horse figure was facing. Then the participant was turned in 180 degrees to face the recall table. On the recall table, the participant had to reproduce the situation from the stimulus table: the line with toys, which had directions. As a result, users of a relative system of spatial orientation put toys on the recall table into directions according to their body, i.e. if some toy on the stimulus table was directed from their left hand to their right hand, then the same toy on the recall table was also directed from their left hand to their right hand. Considering that participants were turned 180 degrees, toys on the recall table were also directed in opposite ways in comparison with toys on the stimulus table. Users of an absolute system of spatial orientation directed toys on the recall table symmetrically to the toys on the stimulus table. In other words, users of different systems of spatial orientation directed toys on the recall table in opposite ways. Therefore the explored Whorfian effect can be described as an influence of language on spatial orientation. More precisely, this Whorfian effect showed that the system of spatial orientation in a language influences the way in which people identify equivalent positions of objects in space.

In the counter experiment by Li and Gleitman only native English speakers participated, but all the participants were divided into three groups. The idea was to carry out the same experiment as was done by Pederson et al., but with different kinds of spatial backgrounds. The first group of participants was examined in an ordinary room with two tables and without any specific background. The second group of participants was examined in a room with open windows and the third group was examined in some open landscape. It turned out that English speakers also used an absolute system of spatial orientation in cases where they were close to some conspicuous landmark (the third group). This means that Li and Gleitman proved that the system of spatial orientation in English language can be flexible and doesn't determine spatial orientation of English speakers. It's

really important to notice that Li and Gleitman changed only the background of the first experiment to reverse the explored Whorfian effect, while the procedure of the experiment remained the same. I think that a change of background can be a feature, which reverses Whorfian effects in a case when a Whorfian effect is sensitive to the background. I will discuss this feature in subsection 2.4. Nevertheless, both experiments by Pederson et al. and by Li & Gleitman demonstrated a really great example of how Whorfian effects can be reversed. But the problem is that all the other Whorfian effects weren't reversed by any counter experiments and Weiskopf & Adams used their own theoretical investigations to prove that other Whorfian effects are also reversible.

2.2. Reversible Whorfian effects in ontology

The experiment by Li & Gleitman demonstrated the only one case, in which the Whorfian effect was reversed practically by a counter experiment. But usually researchers don't have an opportunity to provide counter experiments for every Whorfian effect. In that case researchers should theoretically detect some properties of a Whorfian effect, which make this Whorfian effect reversible and will not allow this Whorfian effect to be repeated in some other experiment. Sometimes, it's really difficult to say without a counter experiment, which property of some Whorfian effect makes it reversible. Weiskopf and Adams recapped several papers, which described differences in ontological categorization by language. One more time I will concentrate on papers, which don't only describe some differences between languages, but also describe some explored Whorfian effects. There are three of these papers: one paper by Lucy (1992) and two papers by Imai and Gentner (1993), (1997).

Imai and Gentner compared Whorfian effects in two groups of participants. The first group of participants was composed of children and the second group was composed of adults. In this experiment, the Whorfian effect was explored in the group of adults, but it was ignored by the group of children in the same experiment. At first glance, the results of the group of children empirically proved that the Whorfian effect is reversible. But it's obvious that all the Whorfian effects can be reversed in childhood. At least infants can't represent any Whorfian effects just because they can't speak at all. Therefore, people obviously should be proficient enough in speaking a language to test them, but then it's not clear whether or not children are proficient enough in speaking a language to test them. Moreover, children of different ages have different levels of language proficiency. That is why it's not clear whether or not there is some age when children become proficient

enough in speaking language to expect that they can be influenced by Whorfian effects. Anyway, I think that this is a really special case of reversibility, which I will discuss in subsection 2.4 with properties of reversible Whorfian effects.

As for the experiment by Lucy, there were two groups of participants: English speakers and Mayan speakers of Yucatec. Yucatec language has an obligatory quantifier of shape for nouns and "two candles" in English can be translated into Yucatec as "two long thin candles" (Weiskopf & Adams 2015, 257). The quantifier in that case is "long thin". In the experiment two groups of participants had to solve a non-linguistic task and choose the two most similar in their opinion candles from the list. The English speaking group matched candles according to their shape, while Yucatec speakers matched candles according to their material. Of course, candles should be always made of wax, but wax can be also different. For example, the used candles were of different color and then their material can be identified as slightly different. After this experiment, Lucy also suggested that in Yucatec language "candle" is translated not by some equivalent word, but by a description with obligatory quantifiers for shape and substance. According to this suggestion, "two candles" in Yucatec can be literally translated into English as "two long thin waxes" (Lucy & Gaskins 2001, 261). That is why Yucatec speakers chose material as the most important feature for match between candles. Weiskopf and Adams concluded that: "When some objects are labeled, it is not clear whether the label they are given attaches to the kind of stuff they are made of or the type of coherent, countable entity they are. The situation is ambiguous. In such unclear cases, the boundary of classification can be *nudged* one way or another by language" (Weiskopf & Adams 2015, 258).

In this quote Weiskopf and Adams wanted to say that the form of labels, i.e. both grammatical and lexical forms of words, sometimes can say something to a speaker. For example, "kohupiim" in Estonian means "quark". "Piim" means "milk" and "kohuma" means "to rise up" after boiling or fermentation. Therefore, the word "kohupiim" can say to Estonian native speakers at least that this product is made of milk. This additional information, which can say something more about the word than any linguistic features (as grammatical gender, grammatical case, grammatical number etc.) names extralinguistic information, i.e. non linguistic information. Kind of the same situation happened with candles in Yucatec language, where "two candles" are named "two long thin waxes". In other words, Weiskopf and Adams noticed that sometimes forms of labels, i.e. forms of words, can say some extralinguistic information about its referent to a native speaker. More precisely, speakers of Yucatec necessarily need to mention material of objects to

name them, which is extralinguistic information. And such extralinguistic information can be used by experimenters as a nudge, i.e. to provoke participants to use this extralinguistic information in their choice. Usually people don't take into account any extralinguistic information. But in the case of a nudge, researchers design their experiments in the way to make participants actualize extralinguistic information. In the experiments by Lucy there was exactly that case, where participants actualized the extralinguistic information to make a choice. The task was to choose the two *most similar* candles, but there were no two identical candles and all the candles were really different. Therefore, participants had only one opportunity to make a considered choice: they had to choose some dominant property of candles and choose two candles as the most similar, if these candles both matched in terms of their dominant property. English speakers chose the two most similar candles according to their form, which is pretty common for a lot of languages. Yucatec speakers chose the two most similar candles according to their material, because their language includes extralinguistic information about a material of objects. In reality people don't use any extralinguistic information and none of English or Yucatec speakers would choose these candles, which they chose in the experiment, as the most similar candles instead of two really identical candles. But in the experiment, participants had no opportunity to choose two really similar candles in their opinion, because they had to choose only from a presented set of candles. Therefore, to use extralinguistic information became the only one possible way to answer given the situation, where the right answer wasn't presented. And Weiskopf & Adams called this strategy of experimenters "nudge". Nudge is obviously a feature, which makes Whorfian effects reversible: no nudge, no Whorfian effect. According to the described features of nudge I can define it as an experimental task, which doesn't include the right answer and provokes participants to answer by using extralinguistic information of their language. For example, the right answer in case of the experiment by Lucy obviously would be two identical candles. But instead of this, participants had to choose between really different candles, which hardly can be called similar. As a result, participants made their choice only by reflection of extralinguistic information.

There is also a question, whether or not nudges in experiments of the Sapir-Whorf hypothesis can be avoided. For example, the system of spatial orientation seems to be a kind of extralinguistic information as well. But the difference is that in the experiment by Pederson et al. a really casual situation was reproduced. People compare objects' positions in space in everyday life very often. Krongauz (2011) mentioned the case that people with

an absolute system of spatial orientation had an opposite opinion from Europeans about *the same* apartment in a hotel. In the hotel were two lines of apartments from both sides of the long corridor and they were located symmetrical to the corridor. Materials and furniture in these apartments were totally the same. That is, if some owner of one apartment sees through the open door of the opposite apartment, then she will see all the layout and furniture like in the mirror. Europeans were absolutely sure that these apartments are identical, but people with absolute system of spatial orientation had the opposite point of view. The thing is that if in one apartment windows opened to the south, then in the symmetrical apartment windows opened to the north. This means that the Whorfian effect explored by Pederson et al. can appear also in everyday life, but at the same time this Whorfian effect was empirically reversed. Therefore, Whorfian effects, which can appear in everyday life, don't include any nudge. While Whorfian effects, which have no chance to appear in everyday life, are most likely a result of a nudge.

2.3. Reversible Whorfian effects in gender

Experiments in gender were presented mostly by Boroditsky et al. (2003) and recapped by Weiskopf and Adams as well, as by Reines and Prinz. The gender argument in the Sapir-Whorf hypothesis is based on the idea that grammatical gender can influence thought. Unfortunately, Reines and Prinz didn't discus somehow reversibility of theWhorfian effects explored by Boroditsky et al. Therefore, this time I will also concentrate on commentaries by Weiskopf and Adams.

In their paper Boroditsky et al. presented a set of experiments and two of them explored a really interesting Whorfian effect. In the first experiment were two groups of participants: native German speakers and native Spanish speakers. All the participants knew also English language and the task was to describe some words in English as if they were real objects, e.g. key, bridge etc. The thing is that German and Spanish languages have grammatical gender and a lot of words in both languages have different genders, e.g. *key* is feminine is Spanish and masculine in German. As a result if some word had the same gender in both languages, then descriptions of such word were also similar in both groups of participants. But if gender of some word differed in languages, then descriptions of the word also differed according to their gender: "The word "key" is masculine in German and feminine in Spanish. German speakers described keys as "hard, heavy, jagged, metal, serrated, and useful", while Spanish speakers described them as "golden, intricate, little, lovely, shiny, and tiny" (Weiskopf & Adams 2015, 262).

This Whorfian effect wasn't reversed by a counter experiment. That is why Weiskopf and Adams tried to prove its reversibility theoretically. They stressed that this experiment didn't demonstrate any strong conclusion about which exact properties of grammatical gender influence thought:

> What sorts of stereotypical properties should we expect to be generated in these languages? This is a perennial problem with attempts to draw large-scale conclusions about the effect of general properties such as gender on thought: the range of linguistic diversity is vastly greater than has been experimentally sampled, and the determinism hypothesis generates no clear predictions for much of the space of this variation. (Weiskopf & Adams 2015, 263)

This means that more precisely not grammatical gender influences thought, but grammatical gender forms some exact stereotypical properties of language, which influence thought. This seems to me to be an attempt to clarify which relations between language and thought were explored by the experiment. This is also really close to what researchers try to do in formulating the scientifically testable version of the Sapir-Whorf hypothesis. But I'm pretty sure that the other experiments recapped by Weiskopf and Adams didn't avoid this problem and the explored relations between language and thought in space and ontological arguments are also not so clearly described. That is why this criticism by Weiskopf and Adams on only gender argument seems really confusing for me. The problem of ambiguous empirical data in the Sapir-Whorf hypothesis should be solved by formulating a scientifically testable thesis. The scientifically testable thesis should prescribe all the properties of empirical data, which make some empirical data capable of supporting the Sapir-Whorf hypothesis. In other words, ambiguity of empirical data one way or another is a problem of every experiment in the Sapir-Whorf hypothesis for today, but ambiguity doesn't make Whorfian effects reversible. Ambiguity of empirical data isn't also the reason, why some experiments should be rejected as evidence for the Sapir-Whorf hypothesis. That is why I don't think that Weiskopf and Adams actually proved that Whorfian effects in gender are reversible. Nevertheless, I think that this Whorfian effect, which was explored in the experiment by Boroditsky et al., can be reversed in another way.

I'm pretty sure that this experiment was also based on the nudge, which was described by Weiskopf and Adams to reject the space argument. In the first experiment by Boroditsky et al. the task was linguistic, i.e. participants had to describe not real objects, but written in English words. Every object has some basic characteristics, which can be described by a respondent from the first look. But if respondents have to describe some object according to the word, then there is a really long list of characteristics which can be

attributed to this object. For example, some key can be silver, gold, long, short, big, small, shiny, rusty, scratched etc. It's pretty common to see more than 100 associations for some word in an associative dictionary. But in the experiment, the task wasn't to remember all the associations, a lot of which are antonyms. The task was to characterize the object according to the given word, i.e. participants had to imagine some concrete object and imagine its concrete properties. One more time the correct answer didn't exist and participants tried to get some hint to answer somehow, but randomly. Grammatical gender became the only one hint in this task and participants oriented their answers on feminine or masculine characteristics. As a result the group of Spanish native speakers described the key by much more feminine characteristics, while the group of German native speakers described the key by much more masculine characteristics. It's also really difficult to imagine that somebody needs to describe some object according to only a word in real life. Therefore, this experiment by Boroditsky et al. obviously included the nudge, which means that the represented Whorfian effect is reversible.

In the second experiment by Boroditsky et al. the same two groups of participants (German and Spanish native speakers) had to *memorize pairs of words*. One word in the pair was some name (male or female) and another word was some proper noun, which meant some object. As a result participants remembered pairs much better, if the sex of the name matched the grammatical gender of the proper noun. Weiskopf and Adams drew the same conclusion about both experiments by Boroditsky et al. that these experiments didn't demonstrate any concrete relations between grammatical gender and thought. I already concluded that this argument can't be used to reject any experiment in the Sapir-Whorf hypothesis. But for the first time I also don't think that the second experiment by Boroditsky et al. included the nudge. I even don't agree with Weiskopf and Adams that this Whorfian effect could be reversed at all, at least so easily as before.

There is no nudge in this experiment, because the situation organized in the experiment can easily happen in everyday life. For example, somebody can organize some celebration: birthday party or wedding. The organizer will make a list with guests and things or presents, which guests should bring. In this case all the notes in the list will look exactly as pairs of words and proper nouns. For example, John – cake, Mary – toaster, Bill – camera etc. Obviously if the organizer has grammatical gender in her language, then she will memorize these pairs of words better, if the sex of the name will match the grammatical gender of the proper noun. Thanks to Boroditsky et al. now we know this.

Nevertheless, I see a problem with this second experiment by Boroditsky et al., but it's not about its reversibility. The thing is that it's difficult to say whether or not grammatical gender is an advantage or disadvantage in memorizing. For example, does grammatical gender boost memorizing for people with grammatical gender in language if the grammatical gender in the pair of words matches the sex like it was in the experiment? In what case should people have the better memorizing in such kinds of tasks: if they don't have grammatical gender in language (therefore grammatical gender is the disadvantage in memorizing if sex and grammatical gender don't match) or if they have grammatical gender in language and sex matches grammatical gender in the task (therefore it's the advantage). The obvious problem is that for such an experiment, two groups of participants seem to be needed: with grammatical gender in language and without. But every person has individual differences in intelligence and it's not clear, how the comparison between two groups of people with different intelligences can be provided. The group of people without grammatical gender in language seems to demonstrate no difference in memorizing of these pairs of words, because there is no difference in grammatical gender between proper nouns. The group of people with grammatical gender in language seems to demonstrate the same results as in the experiment by Boroditsky et al., but it seems incorrectly to compare these two groups of participants.

As a result, the first experiment by Boroditsky et al. obviously included the nudge and it makes this experiment reversible. The second experiment by Boroditsky et al. didn't include the nudge, but provided ambiguous empirical data, which was stressed by Weiskopf and Adams. But this can't be a reason to reject this data as incapable to support the Sapir-Whorf hypothesis. Therefore, I think that the second experiment by Boroditsky et al. provided a really good candidate for irreversible Whorfian effects. But to be sure whether or not this Whorfian effect is irreversible, I should summarize all the properties of reversible Whorfian effects to be sure that the explored by Boroditsky et al. Whorfian effect can avoid all of them.

2.4. Properties of reversible Whorfian effects

In subsection 2.1 I showed that Whorfian effects can be reversed, if the background was changed. Then in subsections 2.2 and 2.3 I showed that a Whorfian effect is reversible, if the nudge was detected in the experiment. In subsection 2.2 I also noticed the feature of Whorfian effects as a tendency to be reversed in childhood. I have explicitly described what the nudge is in subsection 2.2. Therefore, in this subsection I will concentrate on

explaining how background can reverse Whorfian effects and what the tendency to be reversed in childhood means for Whorfian effects.

The nudge was used very often in the experiments of the Sapir-Whorf hypothesis, while the change of background was presented only in the experiment by Li and Gleitman. The thing is that the change of background can obviously reverse only Whorfian effects, which are sensitive for such change. The experiments in the space argument were obviously sensitive to the change of background, because background is a part of space. Background in that case is opposite to foreground. In the experiment by Li and Gleitman foreground was presented by two tables and toys on these tables, while background was presented by the place, where these tables were located (the room or open landscape). Probably background can be presented not only by a place of an experiment, but for now it's difficult to say how else. The nudge is a much more universal feature, which helps to identify a reversible Whorfian effect. Nonetheless, the change of background is the feature, which for sure can reverse Whorfian effects in space. Therefore, the change of background should be also taken into consideration when researchers work with Whorfian effects.

I've already mentioned in subsection 2.2 the tendency of Whorfian effects to be reversed in childhood. In the experiment by Imai and Gentner (1993), (1997) the same Whorfian effect was presented, which was explored in a group of adult participants, but disappeared in a group of children. In subsection 2.2 I also mentioned that all the Whorfian effects actually can be reversed in this way. For example, infants obviously have no Whorfian effects, but I think that in reality nobody expects from infants to demonstrate Whorfian effects. This means that experimenters work only with proficient users of language. Moreover, the mother tongue of participants is also really important in experiments. For example, in the experiments by Boroditsky et al. participants had to be native speakers of German and Spanish. Nobody expected from beginners, who just started to study German or Spanish, to demonstrate the same Whorfian effects as the native speakers. That is, native speaker and non-native speaker are two different levels of language proficiency. Therefore, children of different ages also can have different levels of language proficiency and sometimes this level can be just insufficient to demonstrate some Whorfian effects.

Imai and Gentner tested two and four year old children, but it's a big question, when children get the necessary level of language proficiency to demonstrate Whorfian effects. Children have sensitive periods for language acquisition (Heine, 2008). This

roughly means that children learn different language structures at different ages. Moreover, every structure should be learnt at an appropriate age. Otherwise sensibility will be lost and a child will have difficulties in learning of language. For example, Japanese people, who never studied English, don't discriminate phonemes *la-ra* and *va-ba* (ibid. 156). Children of English speaking parents lose ability to discriminate some phonemes in Hindi language at about one year old (ibid. 156-157). This proves that children study language gradually, as far as they gradually become able to demonstrate Whorfian effects. Therefore, it seems necessary to test Whorfian effects on adults, even if the border between children and adults is vague and probably can vary. Otherwise, it's incorrect to speak about the influence of language on thought, if a tested person doesn't have the necessary level of language proficiency.

I think that the level of language proficiency can include many factors. Probably some Whorfian effects can be ignored by bilinguals or by people, who learned some foreign language in their adulthood, but really perfectly. For example, Henry Kissinger was 15, when his family came to the United States. He spoke English with a German accent even after decades of living in the USA. But his brother Walter was 14 and could get rid of the German accent. Heine used this example to explain how sensitive periods in studying language work (ibid. 156-157). It would be obviously difficult to find a group of people, like Henry or Walter Kissingers, to test whether or not a foreign language influenced their Whorfian effects. Immigrants come to different countries at different ages and also have different motivations and different speeds in studying language. Nevertheless, I agree that that probably people can get rid of some Whorfian effects by intensive study of some specific foreign language. For example, somebody can be even a bilingual speaker of English and let's say German, but if some Whorfian effect is common for both languages, then nothing will change. At the same time, if such a person studies Yucatec language or is a bilingual speaker of English and Yucatec, then Whorfian effect from the experiment by Lucy probably can be reversed.

Therefore, I'm sure that experimenters should test whether or not Whorfian effects are reversible only with appropriate participants. Characteristics of appropriate participants for tests of Whorfian effects probably should be defined by separate research. But for now it's obvious that appropriate participants should have some certain language skills. It can be necessary for participants to be native or fluent speakers of some language or languages. Sometimes, it can be conversely important for participants to have no skills in some language. Appropriate participants also should be adults and lack any mental diseases. Of

course, for some certain tests exactly children, bilinguals, people with some diseases or some other unusual groups of people can be needed. But in that case it will be unfair to claim that some Whorfian effect was reversed. That is why the experiment by Imai and Gentner doesn't really show that the tested Whorfian effect was reversed. This experiment only shows that such a special group of language users as children isn't influenced by Whorfian effects.

Summarizing this subsection, Whorfian effects can be reversed by the change of background, by detecting the nudge and also experiments on Whorfian effects should be provided by appropriate participants. The term "appropriate participants" can seem disputable. Nevertheless, I'm sure that my discussion on appropriate participants showed that some really special groups of people, like children, don't need to be taken into consideration in the research on reversible Whorfian effects.

Conclusion

In this chapter I defined the properties of reversible Whorfian effects. These properties should help in future research of the Sapir-Whorf hypothesis and in formulating of its scientifically testable thesis. Firstly, these properties should help researchers to identify reversible Whorfian effects faster and more precisely. Secondly, these properties will help me to identify irreversible Whorfian effects in chapter three. That is, if Whorfian effects lack properties of reversible Whorfian effects, then these Whorfian effects are good candidates for identification as irreversible Whorfian effects. And thirdly, the properties of reversible Whorfian effects will help me to show in chapter two that sometimes researchers mix up reversibility with inability of empirical evidence to support the Sapir-Whorf hypothesis. For example, Weiskopf and Adams have proven during the whole paper that all the presented Whorfian effects were reversible. But the fact that these Whorfian effects are reversible doesn't make them unable to support the Sapir-Whorf hypothesis. I will show this straight away in the third chapter.

3. The Sapir-Whorf hypothesis. Terms and interpretations

I have two goals in this chapter. My first goal is to prove that Reines & Prinz and Weiskopf & Adams tested the same interpretation of the Sapir-Whorf hypothesis and identified the Whorfian effects represented in their papers as reversible. Otherwise, my comparison of their papers would make no sense. The problem is that Reines & Prinz and Weiskopf & Adams used totally different terminology and explanations. Sometimes it's really difficult to say whether or not authors really were talking about the same thing in their papers. In other words Reines & Prinz and Weiskopf & Adams were arguing in their papers whether or not reversible Whorfian effects can support the weak form of the Sapir-Whorf hypothesis, but this should be justified. Therefore, in subsection 3.1 I will clarify and compare all the terminology, which was used by Reines & Prinz and Weiskopf & Adams. In subsections 3.2 and 3.3 I will recap interpretations of the Sapir-Whorf hypothesis, which were discussed by Reines & Prinz and Weiskopf & Adams. I need to do this because interpretations of the Sapir-Whorf hypothesis are much more complicated than ordinary terms and all the necessary explanations take a lot of space.

My second goal is to show that without a clear definition of reversibility and reversible Whorfian effects, researchers can misinterpret empirical data in the Sapir-Whorf hypothesis. Reines & Prinz are researchers, who firstly noticed that "Whorfian effects are often reversible" and briefly described these effects. But the real goal of their paper was to stress that reversible Whorfian effects can perfectly support some interpretations of the Sapir-Whorf hypothesis. And this is a good example of how reversible Whorfian effects should be researched. At the same time, Weiskopf and Adams stressed that reversible Whorfian effects can't support the Sapir-Whorf hypothesis. But in reality, Weiskopf and Adams only showed that the experiments recapped in their paper were reversible, but this doesn't mean that reversible Whorfian effects can't support the Sapir-Whorf hypothesis. Reines and Prinz defined reversibility as a common property of Whorfian effects, while Weiskopf and Adams seemed to explain a new reason to reject each experiment. As a result, Weiskopf and Adams provided really good material for defining reversibility and its properties, but didn't justify enough why they rejected these reversible Whorfian effects. That is why I will also add the explanations by Reines & Prinz and Weiskopf & Adams of why they rejected or accepted some interpretation of the Sapir-Whorf hypothesis to be supported by empirical data.

As a result, this terminological verification will reduce the risk that I am misinterpreting Reines & Prinz and Weiskopf & Adams. My recap of interpretations by Reines & Prinz and Weiskopf & Adams will also show that the research of reversibility is really important in the dispute on whether or not empirical data can support the Sapir-Whorf hypothesis.

3.1. Terms and concepts of the Sapir-Whorf hypothesis

The Sapir-Whorf hypothesis by itself can be called a lot of different ways: *the linguistic relativity hypothesis, the principle of linguistic relativity, Whorfianism, linguistic determinism,* and I'm pretty sure that this is far away from the full list of its possible names. I agree with Krongauz, who said that "I prefer […] the title "the Sapir-Whorf hypothesis" in honor of two remarkable scientists" (Krongauz 2012). Therefore, I will call this hypothesis *the Sapir-Whorf hypothesis*. But when I recap somebody's interpretation I will use their titles, i.e. *Whorfianism* by Reines & Prinz and *linguistic determinism* by Weiskopf & Adams. The Sapir-Whorf hypothesis also has several interpretations and this is a problem for researchers, who try to find out whether or not the Sapir-Whorf hypothesis can be supported by empirical data. Roughly, the Sapir-Whorf hypothesis means that *language somehow affects thought*, but its different interpretations describe the influence of language on thought differently, e.g. that language *determines* thought (the strong form) or *influences* thought (the weak form) (ibid. 2012, 00:00 – 02:00), (Ahearn 2011, 69). As a result, different interpretations make different demands on empirical data, which should support them. That is why different researchers can have different opinions on whether or not some empirical data can support some interpretation of the Sapir-Whorf hypothesis. But in reality, neither the strong nor the weak forms of the Sapir-Whorf hypothesis are clear enough to be tested by empirical data. This problem is obviously addressed to the formulation of the scientifically testable thesis as I stressed in the introduction. And while the scientifically testable thesis is far away from formulation, researchers of the Sapir-Whorf hypothesis work with its existing interpretations and test whether or not empirical data can support them.

Reines & Prinz and Weiskopf & Adams didn't define what they meant by *empirical data* and obviously used this term as an intuitively clear concept, but I prefer to specify it in my thesis for better understanding of what Whorfian effects and reversibility are. By *empirical data*, I mean experimental results. Reines & Prinz and Weiskopf & Adams also used terms *empirical data* and *experimental results* as synonyms. Usually every

experiment in the Sapir-Whorf hypothesis includes *stimulus* and *reaction*. Stimulus can be presented by linguistic or non-linguistic tasks. A linguistic task means that participants of the experiment had to deal with some language stuff, e.g. to remember some words or to read some word and describe its referent etc. A non-linguistic task means that language wasn't involved in the assignment and participants had to deal with some objects or colors etc., e.g. to choose the two most similar colors, to choose the two most similar objects, to choose the brighter color etc. Reaction is presented by answers of participants to experimental task.

In addition to stimulus and reaction, empirical data always include some interpretation of results, i.e. researchers try to find out what empirical data can say about relations between language and thought. As I mentioned in the introduction, authors of psycholinguistic experiments don't have the goal of proving or disproving the Sapir-Whorf hypothesis. But at the same time, authors of the psycholinguistic experiments always interpret explored results as concrete relations between some properties of language and some properties of thought, e.g. as influence of grammatical gender on memorizing, influence of color categorization in language on color discrimination etc. Then some scientists, like Reines & Prinz and Weiskopf & Adams, try to find out whether or not the explored relations between some properties of language and some properties of thought are enough to support the Sapir-Whorf hypothesis and claim that in that case, language determines or influences thought. These experimentally explored relations between some properties of language and some properties of thought were called by Reines & Prinz *Whorfian effects*. Weiskopf & Adams didn't use any common term, but their *determining effect*, *linguistic effect* or just *effect* have obviously the same meaning according to context.

Reversibility is an ability of Whorfian effects to be reversible. In general, if Whorfian effects are reversible, then they don't influence any cognitive abilities anywhere but in the narrow experimental situation. In chapter one I've already introduced how Whorfian effects can be reversed practically (in the experiment by Li and Gleitman) or theoretically (in the experiments by Boroditsky et al. and by Lucy). Weiskopf and Adams didn't use any special term for reversibility as well. At the same time, I showed in chapter two that they reversed all the Whorfian effects to stress that these Whorfian effects can't support any interpretation of the Sapir-Whorf hypothesis. One time, Weiskopf & Adams and Reines & Prinz used the same couple of experiments (by Pederson et al. and by Li & Gleitman). Reines and Prinz called the Whorfian effect presented by the experiments reversible, while Weiskopf and Adams concluded that these experiments in spatial

orientation lack “a strong determining effect” (Weiskopf & Adams 2015, 261). By paraphrasing Weiskopf and Adams, this was *a weak determining effect*, which was called by Reines and Prinz reversible Whorfian effect. Then the strong determining effects should be presented by *irreversible Whorfian effects*. But neither Weiskopf & Adams nor Reines & Prinz used the term “irreversible Whorfian effects”. Though Weiskopf and Adams mentioned the strong determining effect (what should be the same as irreversible Whorfian effect), they didn’t define its properties and even didn’t assume whether or not this strong determining effect exists. In chapter four, I will introduce one Whorfian effect, which seems to avoid properties of reversible Whorfian effects. Probably such Whorfian effects can be called irreversible.

That is all that I wanted to say about terms in papers by Weiskopf & Adams and Reines & Prinz. But these defined terms mostly should help me to explain in subsections 3.2 and 3.3 why Reines & Prinz stressed that reversible Whorfian effects can support the Sapir-Whorf hypothesis, while Weiskopf & Adams insisted on rejecting these Whorfian effects as empirical evidence for the Sapir-Whorf hypothesis.

3.2. Interpretations by Reines & Prinz

In this chapter, I will recap interpretations of the Sapir-Whorf hypothesis by Reines and Prinz to be sure that I understood correctly what they were doing in their paper and show how researchers should deal with reversible Whorfian effects. Reines and Prinz detected reversible Whorfian effects and developed the already existing weak form of the Sapir-Whorf hypothesis to prove that reversible Whorfian effects can support the Sapir-Whorf hypothesis. As a result, their Habitual Whorfianism and Ontological Whorfianism became a good attempt in formulating of the scientifically testable thesis of the Sapir-Whorf hypothesis, which can be tested by empirical data. Nevertheless, Reines and Prinz didn’t mention whether or not irreversible Whorfian effects exist and what their place in the Sapir-Whorf hypothesis is. In chapter three, I will introduce my point of view on whether or not irreversible Whorfian effects exist, what can become a good topic for formulating of the scientifically testable thesis of the Sapir-Whorf hypothesis. In the current subsection I also should show that I understood correctly how Reines and Prinz described Habitual Whorfianism, Ontological Whorfianism and reversible Whorfian effects, because their explanations sometimes are brief and sometimes lack necessary definitions.

I’ve already mentioned that the Sapir-Whorf hypothesis have strong and the weak forms. Weiskopf and Adams have discussed three interpretations of the Sapir-Whorf

hypothesis in their paper, while Reines and Prinz discussed even four of them. Since in both papers there are more than two interpretations of the Sapir-Whorf hypothesis, this means that at least some of these interpretations differ from the classical strong and the weak forms. In some cases, it can be described by the lack of unified terminology. For example, I will show a little bit later that *Radical Whorfianism* by Reines & Prinz and strong linguistic determinism by Weiskopf & Adams are obviously the same as the classical strong form of the Sapir-Whorf hypothesis. But sometimes, researchers of the Sapir-Whorf hypothesis create an original interpretation of the Sapir-Whorf hypothesis fr some reason. Reines and Prinz created their *Habitual Whorfianism* and *Ontological Whorfianism* to stress that reversible Whorfian effects can support the Sapir-Whorf hypothesis: "We introduce two theses that would have important implications if true: Habitual Whorfianism and Ontological Whorfianism. We argue that these offer the most promising interpretations of the emerging evidence" (Reines & Prinz 2009, 1022). By *emerging evidence,* Reines and Prinz obviously meant empirical data. Reines and Prinz didn't write somewhere specifically that reversible Whorfian effects can support the Sapir-Whorf hypothesis, but they recapped reversible Whorfian effects (e.g. explored by Pederson et al. and Boroditsky et al.) as able to support Habitual Whorfianism or Ontological Whorfianism. Therefore, I think it's correct to say that according to Reines and Prinz, reversible Whorfian effects can support Habitual Whorfianism or Ontological Whorfianism.

Reines and Prinz represented four interpretations of the Sapir-Whorf hypothesis in total: *Trivial Whorfianism*, *Radical Whorfianism*, *Habitual Whorfianism* and *Ontological Whorfianism*. First of all, Reines and Prinz rejected Trivial Whorfianism and Radical Whorfianism as incapable of supporting the Sapir-Whorf hypothesis, therefore I prefer to start with them. According to Reines and Prinz, *Radical Whorfianism* means that: "languages influence psychological processes because thinking depends on natural language" (Reines and Prinz 2009, 1027). The part of their claim *thinking depends on natural language* is really close to the classical thesis of the strong form of the Sapir-Whorf hypothesis that *language determines thought*. Both these claims basically mean that *thinking is impossible without language*. Reines and Prinz rejected this thesis as incapable to be supported by empirical data because of strong counterarguments:

> The idea that all thought depends on language strikes us as completely implausible. We know from research on mental imagery (Kosslyn et al. 2006), language-deprived adults (Schaller 1991), transient aphasia (Lecours and Joanette 1980), and animal cognition (Hauser 2000), that

> sophisticated decision-making can be achieved without language. There is also a principled argument against Radical Whorfianism form language learning (Fodor 1975). (Reines & Prinz, 2009, P. 1027)

Many researchers of the Sapir-Whorf hypothesis also noticed that the strong form of the Sapir-Whorf hypothesis is too strong to be supported by empirical data (Ahearn 2011, 69). Later I will show that Weiskopf and Adams also rejected the strong form.

According to Reines and Prinz, *Trivial Whorfianism* means that: "[…] languages influence psychological processes because, when we use words, we draw attention to things that we might happen to neglect without it" (Reines & Prinz 2009, 1028). Reines and Prinz also noticed that: "This is trivial because no one doubts that language can direct attention. If I say, 'Look up in the sky!' you may follow this command and notice something that would have otherwise gone unseen" (ibid. 1028). But the problem is that Trivial Whorfianism doesn't need any empirical data to be proven. Reines and Prinz had a goal in their paper to support by empirical data the most informative interpretation of the Sapir-Whorf hypothesis, which can say about relations between language and thought as much as possible. Therefore, Trivial Whorfianism was also rejected, because empirical data can't help in that case to learn something new about language and thought.

By *Habitual Whorfianism* Reines and Prinz meant that: "languages influence psychological processes because they instill habits of thought that lead us to think in certain ways by default that we would not have thought in without language learning" (ibid. 1028). As written above, Whorfian effects are relations between some properties of language and some properties of thought. Reines and Prinz didn't define what these *habits of thought* are, but I think that by *habits of thought* Reines and Prinz mentioned exactly Whorfian effects. The reason is that Whorfian effects also seem to lead us to think in certain ways, as Reines and Prinz wrote about habits of thought. Otherwise, whether habits of thought and Whorfian effects are different things, I can only guess that habits of thought cause Whorfian effects. But in that case, it's totally unclear what forms these habits of thought. For example, grammatical gender influences memorizing of some words, like it was explored in the experiment by Boroditsky et al. In that case, grammatical gender is a cause, while some appeared features in memorizing are a consequence. All together the cause and the consequence are the Whorfian effect. How can habits of thought influence Whorfian effects then? Obviously habits of thought can't influence grammatical gender or the way how grammatical gender influences memorizing. Therefore, I think that habits of thought are just Whorfian effects of a certain kind. As an example of such a habit of

thought, Reines and Prinz used the experiments on grammatical gender by Boroditsky et al. The other type of Whorfian effects are described by Ontological Whorfianism.

I can also guess that by this term *habits of thought* Reines and Prinz wanted to stress that people aren't determined to be influenced by some Whorfian effect, but have a habit to be influenced, e.g. by grammatical gender. One can object that habits are a much more regular thing than Whorfian effects, especially than reversible Whorfian effects. A smoking person smokes several times a day, while nobody knows exactly how often Whorfian effects appear in real life. But habits also can appear really rarely. For example, somebody has a habit to go skiing on weekends. If there is little snow in the winter, like it often happens in Estonia, then such a person can go skiing only a couple of times during the whole year. Therefore, I think that even reversible Whorfian effects can be called habits of thought.

According to Reines and Prinz *Ontological Whorfianism* means that: "languages influence psychological processes because they lead us to organize the world into categories that differ from those we would discover without language" (ibid. 1029). Reines and Prinz also noticed that usage of these categories is a kind of habits of thought: "In leading us to habitually group certain particulars together (an effect of Habitual Whorfianism), language shapes the categorical boundaries that constitute our subjective organization of world" (ibid. 1029). Then Ontological Whorfianism can be described as a variant of Habitual Whorfianism, but this time *ontological* habits of thought influence thought: "[…] language influences our understanding of what kinds of things exist – our ontologies" (ibid. 1029). Ontologies are obviously presented by Whorfian effects of certain kind, e.g. the experiments in the ontological argument.

Summarizing features of Habitual Whorfianism and Ontological Whorfianism, I think that they can be roughly paraphrased as *language influences thought*, which is the thesis of the weak form of the Sapir-Whorf hypothesis. The difference is that Reines and Prinz clarified this thesis a lot in their interpretations. In forms of Habitual Whorfianism and Ontological Whorfianism the Sapir-Whorf hypothesis prescribes demands for empirical data really clearly. Habitual Whorfianism and Ontological Whorfianism define which properties empirical data should have and which relations between language and thought such empirical data can prove. This seems to be a good try to formulate the scientifically testable thesis. In the next subsection, I will show that Weiskopf and Adams

conversely didn't improve the weak form of the Sapir-Whorf hypothesis, but still tried to test whether or not this can be supported by empirical data.

As a result, Reines and Prinz concluded that: "Philosophers who say that language is merely a vehicle for expressing thoughts must revise their views, and it may turn out that speakers of different languages habitually parse the world in different ways" (ibid. 1030). In other words, Reines and Prinz think that they successfully proved that the Sapir-Whorf hypothesis in face of Habitual Whorfianism and Ontological Whorfianism is supported by empirical data, including reversible Whorfian effects. To justify this claim, Reines and Prinz stressed that habits of thought and ontologies are presented by recapped Whorfian effects, which are "often reversible". In that case, the Sapir-Whorf hypothesis was interpreted as influence of habits of thought on thought and as influence of ontologies (aka ontological habits of thought) on thought. I think that this was a really big step in formulating the scientifically testable thesis. Nevertheless, the problem is that Reines and Prinz didn't mention, whether or not irreversible Whorfian effects exist. Probably, irreversible Whorfian effects can become a base for some new interpretations of the Sapir-Whorf hypothesis, which can say more about relations between language and thought. At the same time, not all researchers agree that reversible Whorfian effects can support the Sapir-Whorf hypothesis and this also a purpose to research reversibility deeper.

3.3. Interpretations by Weiskopf and Adams

In this subsection I will recap interpretation by Weiskopf and Adams to show that I understood what they were doing in their paper correctly and that sometimes researchers mix up reversibility with the reason to reject empirical data as evidence for the Sapir-Whorf hypothesis. Weiskopf and Adams only proved in their paper that all the presented Whorfian effects were reversible, but they didn't prove that reversible Whorfian effects can't support the Sapir-Whorf hypothesis. As a result, Weiskopf and Adams rejected reversible Whorfian effects as empirical evidence for the Sapir-Whorf hypothesis without necessary justification. Moreover, for some reason Weiskopf and Adams didn't touch these Whorfian effects, which can be irreversible in my opinion. Therefore, their rejecting of reversible Whorfian effects as empirical evidence for the Sapir-Whorf hypothesis wasn't reasonable. In this subsection, I will also show that I understood correctly what Weiskopf and Adams were doing in their paper and that they really rejected reversible Whorfian effects as evidence for the weak form of the Sapir-Whorf hypothesis.

Weiskopf and Adams called the Sapir-Whorf hypothesis *linguistic determinism*. They named interpretations of the Sapir-Whorf hypothesis *strong linguistic determinism, ultrastrong linguistic determinism* and *weak linguistic determinism*. Weiskopf and Adams defined the Sapir-Whorf hypothesis exclusively as the strong form of linguistic determinism: "[…] the thesis of linguistic determinism, particularly in its strong form, is sometimes called the Sapir-Whorf hypothesis" (Weiskopf & Adams 2015, 251). I think that the term "linguistic determinism" can be easily associated with the strong form of the Sapir-Whorf hypothesis, because "determinism" can be understood in the way that language *determines* thought, which is actually the strong form of the Sapir-Whorf hypothesis. The terms *strong* and *weak linguistic determinism* aren't commonly used by researchers of the Sapir-Whorf hypothesis anyway. Therefore, I will use these terms only to refer to interpretations by Weiskopf and Adams. In other cases, I will use the term *the Sapir-Whorf hypothesis* to name the hypothesis of linguistic relativity in general and also terms "the weak form" and "the strong form". In subsection 3.1 I explained why I use exactly this terminology.

Weak linguistic determinism by Weiskopf and Adams is obviously the weak form of the Sapir-Whorf hypothesis and I will show it a little bit later. Weiskopf and Adams used exactly weak linguistic determinism for tests by empirical data. Strong linguistic determinism and ultrastrong linguistic determinism are obviously variations of the strong form of the Sapir-Whorf hypothesis, which were rejected by Weiskopf and Adams as incapable to be supported by empirical data. According to Weiskopf and Adams ultrastrong linguistic determinism is:

> […] the claim that language is needed for the very existence of thought. There is no such thing as thinking without some language, either because language is the seed from which higher thought grows, or because language is the inner medium that constitutes such thought. (ibid. 250)

The ordinary strong determinism Weiskopf and Adams defined by saying: "that certain types of thought are only possible given the possession of a certain type of language" (ibid. 250). In reality, Weiskopf and Adams don't seem to distinguish between their strong linguistic determinism and ultrastrong linguistic determinism, because they used the same arguments for both: "Arguments in favor of the strong and ultrastrong theses have been given by philosophers such as W. V. Quine (1960), Jonathan Bennett (1988), and Donald Davidson (1975/1984, 1982/2001)" (Weiskopf & Adams 2015, 253). As a result, strong linguistic determinism and ultrastrong linguistic determinism were distinguished only by different definitions, but I still don't see how two different claims can be supported by the

same argument. Nevertheless, both strong linguistic determinism and ultrastrong linguistic determinism were rejected by Weiskopf and Adams as incapable to be supported by empirical data. That is why differences between strong linguistic determinism and ultrastrong linguistic determinism are irrelevant for research of reversible Whorfian effects.

Step by step Weiskopf and Adams rejected arguments by all these philosophers. Firstly, Weiskopf and Adams rejected Quine's argument (1960): "Quine's restriction of manifestability may be more about limiting languageless creatures to simple thoughts rather than a claim that no thought can be possessed by them" (Weiskopf & Adams, 2015, 250). Secondly, they rejected the argument by Benett (1964/1989; 1988): "In a similar spirit, Jonathan Bennett argued not that there could not be thought without language, but that there could not be thoughts of certain types without language" (Weiskopf & Adams 2015, 253). And finally, Weiskopf and Adams rejected Davidson's argument (1975/1984), (1982/2001):

> Where Bennett is skeptical that animals or languageless creatures could have thoughts of certain kinds without language, Donald Davidson questions whether they could have thoughts at all. […] One does not need the concept of cancer to acquire cancer, or the concept of thought to have a thought. No strong conclusions about the nature of thought itself follow from the epistemic conditions under which we attribute thoughts. (Weiskopf & Adams 2015, 253-256)

In other words, Weiskopf and Adams rejected strong linguistic determinism and ultrastrong linguistic determinism as unfit to be supported at all, no matter with or without the empirical data. This point of view is pretty common and the strong form of the Sapir-Whorf hypothesis is usually rejected by researchers, who try to test the Sapir-Whorf hypothesis by empirical data (Reines & Prinz 2009, 1022), (Pöhls 2013, 101), (Ahearn 2011, 69).

According to Weiskopf and Adams, *weak linguistic determinism* "[…] holds that language influences thought in some way. Most frequently, weak determinism involves the claim that language affects "habitual thought" (Whorf's term) by biasing attention, memory or preferences (Weiskopf & Adams 2015, 251). Weak linguistic determinism seems to be the same as the weak form of the Sapir-Whorf hypothesis, because weak linguistic determinism repeats the claim "language influences thought". At the same time Weiskopf and Adams didn't specify clearly, what *habitual thought* means. This term seems to be really close to *habits of thought* by Reines and Prinz. Nevertheless, Weiskopf and Adams didn't mention how habitual thought works. That is why I can't say that weak

linguistic determinism is something more than the ordinary weak form of the Sapir-Whorf hypothesis. Weiskopf and Adams divided weak linguistic determinism into three groups: *ontology*, *space* and *gender*. I've already presented Whorfian effects from these arguments in chapter one. I was concentrated mostly on how Weiskopf and Adams proved that Whorfian effects were reversible. Now I want to show also why Weiskopf and Adams concluded that reversible Whorfian effects can't support the Sapir-Whorf hypothesis.

Weiskopf and Adams rejected the ontological argument by saying that: "Speaking English may encourage people to conceptualize potentially ambiguous entities as objects, whereas speaking Japanese may encourage conceptualizing them as substances. This type of influence is consistent with, at most, a form of weak determinism" (ibid. 258). The last sentence in this quote seems to give a chance to weak linguistic determinism. I think that Weiskopf and Adams just want to be careful and don't make predictions about future research. But for today they definitely didn't approve Whorfian effects in ontological argument as empirical evidence for the Sapir-Whorf hypothesis because these Whorfian effects are reversible.

The space argument was also rejected by Weiskopf and Adams: "These results, like all others, are suggestive rather than definitive. […] Perception and thought about space seem only weakly Whorfian" (ibid. 261). This time Weiskopf and Adams are also very careful in their claims. Nevertheless, Weiskopf and Adams noticed that: "they [Whorfian effects] are consistent with a view on which languages may encode many different spatial relations and reference frames without these encodings exerting a strong determining effect on spatial cognition itself" (ibid. 261). In other words, Weiskopf and Adams concluded that the strong determining effect is needed to say precisely whether or not the Sapir-Whorf hypothesis can be supported by Whorfian effects in the spatial arguments. Moreover, I've already mentioned in subsection 3.1 that the strong determining effect is the same as irreversible Whorfian effect. Anyway, the conclusion by Weiskopf and Adams about the space argument in the Sapir-Whorf hypothesis doesn't allow to think that they approved Whorfian effects in space argument as empirical evidence for the Sapir-Whorf hypothesis.

Weiskopf and Adams rejected the gender argument because:

> This is a perennial problem with attempts to draw large-scale conclusions about the effect of general properties such as gender on thought: the range of 1inguistic diversity is vastly greater than has been

> experimentally sampled, and the determinism hypothesis generates no clear predictions for much of the space of this variation. (Ibid. 263)

One more time Weiskopf and Adams are careful in their predictions, but they obviously didn't approve Whorfian effects in the gender argument as well.

As a result Weiskopf and Adams showed that the presented Whorfian effects are reversible and made really vague conclusions about them. These conclusions don't let us say straightly that Weiskopf and Adams rejected all these Whorfian effects as empirical evidence for the Sapir-Whorf hypothesis, but at the same time, Weiskopf and Adams obviously didn't approve the presented Whorfian effects either. It's not the same thing to reject Whorfian effects and not to approve these Whorfian effects to be empirical evidence for the Sapir-Whorf hypothesis. In other words, Weiskopf and Adams let an opportunity that some additional research will prove that the presented by Weiskopf and Adams Whorfian effects really support the Sapir-Whorf hypothesis. Therefore, the conclusion by Weiskopf and Adams can be interpreted in the following way: for today there's no evidence that reversible Whorfian effects can support the Sapir-Whorf hypothesis. But in reality, Weiskopf and Adams made a mistake, which I already mentioned: they proved that the presented Whorfian effects are reversible, but didn't prove that reversible Whorfian effects can't support the Sapir-Whorf hypothesis.

Conclusion

In this chapter, I justified my claim that Reines & Prinz and Weiskopf & Adams really discussed the same interpretation of the Sapir-Whorf hypothesis and whether or not reversible Whorfian effects can support this interpretation. Reines & Prinz slightly clarified the weak form of the Sapir-Whorf hypothesis to deal with empirical data and especially with reversible Whorfian effects, while Weiskopf & Adams left this without changes. As a result, Weiskopf & Adams made a mistake: they rejected all the Whorfian effects just because they were reversible without any additional justification. My research on reversibility should help to avoid such misunderstanding in future research. This chapter also showed that in future research on empirical data in the Sapir-Whorf hypothesis it would be really useful to find out whether or not reversible Whorfian effects can support the Sapir-Whorf hypothesis.

4. Irreversible Whorfian effects

In this chapter I will introduce my candidate to the irreversible Whorfian effect, define its properties and explain, why I think this is irreversible. I've already mentioned that the second experiment by Boroditsky et al. (2003) explored a good candidate to be called irreversible Whorfian effect. But the Whorfian effect, which I will introduce in this chapter, seems to be much more convincing. This Whorfian effect was presented by Deutscher (2010), who had used a set of experiments by Winawer et al. (2007), Gilbert et al. (2006) and Tan et al. (2008) to defend the statement that the Sapir-Whorf hypothesis is strongly supported by empirical data. Unfortunately, Deutscher didn't mention whether or not this Whorfian effect can't be reversed. Deutscher even didn't justify somehow that these experiments really support some interpretation of the Sapir-Whorf hypothesis, but stressed that every experiment in the set complements each other and they all together form strong evidence for the Sapir-Whorf hypothesis. For some reason, Weiskopf and Adams didn't mention any of these experiments. Reines and Prinze mentioned only the experiment by Winawer et al. in passing as a part of the ontological argument, but didn't specify whether or not it's reversible or irreversible.

I will recap the set of experiments, which demonstrated my candidate for irreversible Whorfian effect, in subsection 4.1. In subsection 4.2 I will check, whether or not this Whorfian effect lacks properties of reversible Whorfian effects. If this Whorfian effect lacks properties of irreversible Whorfian effect, then this doesn't make this Whorfian effect irreversible automatically. Therefore, in subsection 4.2 I will also define properties of this Whorfian effect, which allow consideration that this Whorfian effect is irreversible.

4.1. The argument by Deutscher

In this subsection I will recap the set of experiments, which were presented by Deutscher as strong evidence for the Sapir-Whorf hypothesis. I think that all these experiments explored the great candidate for the irreversible Whorfian effect. Deutscher noticed that three experiments by Winawer et al. (2007), Gilbert et al. (2006) and Tan et al. (2008) explored the same Whorfian effect. All the experiments tested an ability to compare colors, i.e. to choose the necessary color from two presented. If this is true, then it's the first time, when the same Whorfian effect was repeated in really different experiments. This already sounds like a good argument to call this Whorfian effect irreversible. But firstly I should prove that this was really the same Whorfian effect in all the experiments.

Winawer et al. provided an experiment also known as Russian blues, where they compared two groups of participants: Russian and English native speakers. Participants had to solve the non-linguistic task and choose a lighter shade from two shades of blue. All the time participants had the same speed of reaction in both groups, but in the case of border shades, Russian native speakers spent a little bit more time. Winawer et al. proposed to explain it as an influence of Russian color terms. The thing is that in English language dark blue and light blue are different shades of the same *basic color* (blue). Basic colors have four distinctive features: it should have only one morpheme (blue-green isn't suitable), it can be used in an everyday situation (ultramarine isn't suitable), it can be used to name all the things (palomino isn't suitable), and it can't be a part of another basic color (scarlet isn't suitable) (Krongauz 2001, 90). But in Russian language, there are two separate basic colors: *goluboy* (light blue) and *siniy* (dark blue). That is why Winawer et al. concluded that Russians were choosing not just between lighter and darker shades of the same color, but between two separate basic colors. Therefore, they spend more time to decide if some border shade matches one color or another. Winawer et al. assumed that participants of Russian blues spend extra time to get feedback from their language, because language presumably defines borders of colors.

This idea that comparison of border colors can take more time was confirmed by Gilbert et al., who tested the idea whether the left hemisphere influences thought or perception. The researcher assumed so because Broca's area is located in the left hemisphere, which has a speech producing function. The left hemisphere processes information from the right visual field of our vision and vice versa. Therefore, Gilbert et al. proposed that the left hemisphere can influence the right visual field because of Broca's area. Gilbert et al. used the same idea as in Russian blues, i.e. to check choice making process between two border colors. But this time only the English speaking participants were tested. In this experiment, participants saw on the screen a symmetrical circle made of 12 small green squares. Sometimes, instead of one of the green squares, a blue one would appear in a random place, and participants had to choose whether it appeared from the right or from the left side. And as in the case with Russian blues, participants had a delay in their choice, but only in the right visual field. In other words, if the green square appeared in the left visual field, then participants reacted faster. But if the green square appeared in the right visual field, then participants reacted slower, than it was with the left visual field. Gilbert et al. assumed that the difference between two visual fields in speed of

choice making was caused by influence of the left hemisphere, i.e. the left hemisphere influenced the right visual field and caused a delay, like it was in Russian blues.

Tan et al. tested by Magnetic resonance imaging (henceforth MRI) whether the color discrimination process activates some brain activity, especially in the left hemisphere. Participants of this experiment had two tasks: one non-linguistic and one linguistic. In both cases, participants saw some stimulus colors on the screen. In the first task, participants saw two colored squares on the screen and if these colors were the same, participants should press a button. The task was totally non-linguistic. Nevertheless, the MRI detected that some areas of the left hemisphere were activated by slightly increased blood pressure. In the second task, participants saw the stimulus color name on the screen and had to call it aloud, while the MRI was scanning. The results showed that the same areas in the left hemisphere were heavily activated. Tan et al. drew the conclusion that the purpose of these areas is to map, where color terms are located even in case a person is just thinking about these colors and doesn't name them. In other words when participants only thought about colors, they also used their linguistic capacities. Deutscher stressed that this conclusion is obviously supported in the experiments by Winawer et al. (2007), Gilbert et al. (2006):

> […] it becomes clear that when the brain has to decide whether two colors look the same or not, the circuits responsible for visual perception ask the language circuits for help in making the decision, even if no speaking is involved. So for the first time, there is now direct neurophysiologic evidence that areas of the brain that are specifically responsible for name finding are involved with the processing of purely visual color information. (Deutscher 2010, 230)

Therefore, I also think that this set of experiments describes the same Whorfian Effect, which is a good argument for its reversibility.

4.2. Properties of irreversible Whorfian effects

In this subsection I will check whether or not the Whorfian effect described in subsection 4.1 lacks properties of reversible Whorfian effect. If this Whorfian effect really lacks properties of reversible Whorfian effect, then this still doesn't mean that this Whorfian effect is irreversible. Therefore, in this subsection I will also define properties of this Whorfian effect, which doesn't allow for reversing this.

As I defined in chapter two, Whorfian effects can be reversed by change of background or if a nudge was detected. As for change of background, experiments in color discrimination don't seem to be sensitive to this. I mentioned in subsection 2.1 that

background can probably be presented not only by spatial background, but somehow else. Nevertheless, the experiments by Winawer et al., Gilbert et al. and Tan et al. seem to be too minimalistic to include some additional features, such as background. Participants in these experiments had to choose one from two proposed answers, what is the minimal condition for every test. The experiments represented in the second chapter include much more conditions. For example, in the experiment by Lucy there was a list of candles for choice, i.e. participants had to choice not one candle from two proposed, but two candles form the whole list of candles. I also mentioned that the list of candles was picked up by experimenters to influence participants' choice, because experimenters didn't include on this list two identical candles, which would be the only one correct answer in the task to choose the two most similar candles.

In such minimal conditions, there is also no place for nudge. Moreover, nudge was often detected in the experiments, where participants had to solve the task without the correct answer, like in the experiment by Lucy. In the experiments by Winawer et al., Gilbert et al. and Tan et al. the correct answer was always presented. In the experiment by Winawer et al. participants had to choose a brighter shade, while the category of brightness is universal for every language (Berlin & Kay 1969). In the experiment by Gilbert et al. participants had the task to choose a green square from 12 blue squares on the screen. And in the experiment by Tan et al. participants had to push the button, when two squares of the same color appeared on the screen. In this experiment, there was a correct answer as well. In the previous experiments, the nudge provoked participants to use extralinguistic information, because the correct answer was absent and participants had only one chance to answer deliberately with the help of extralinguistic information. But in these three experiments participants has no need to use extralinguistic information, because the right answer was always presented and participants had no difficulties to find it.

This means that the described by Deutscher Whorfian effect really lacks properties of reversible Whorfian effects. Moreover, its minimalistic design seems to defend this Whorfian effect from any other properties of reversible Whorfian effects, which can be explored in the future. In other words, there is no place in this Whorfian effect for any property of reversible Whorfian effect. At the same time, this Whorfian effect can be obviously reversed in childhood, like all the Whorfian effects, as I mentioned in subsection 2.4. In other words, irreversible Whorfian effects can't be reversible in case of appropriate participants, as I also mentioned about reversible Whorfian effects. This Whorfian effect was also repeated in three different experiments, while previous Whorfian effects lack

repeatability. Deutscher explained this Whorfian effect as a universal process in comparison between different colors, i.e. this Whorfian effect should theoretically appear every time, when people compare two colors.

Conclusion

In this chapter I recapped the Whorfian effect, which was detected by Deutscher in experiments by Winawer et al., Gilbert et al. and Tan et al. This Whorfian effect really lacks properties of reversible Whorfian effect, i.e. this Whorfian effect isn't sensitive to change of the background and doesn't include nudge. This Whorfian effect also includes such unique properties as minimalistic design and repeatability. But the strongest argument for irreversibility of this Whorfian effect was made by Deutscher, who claimed that this Whorfian effect appears every time when people compare two colors. This claim was based mostly on the experiment by Tan et al. Unfortunately, the experiment by Tan et al. doesn't exclude the possibility that the explored brain activation can be caused by some other mental function instead of the described influence of language. Moreover, even if this brain activation was really caused by language influence, then the experiment by Tan et al. also doesn't exclude a possibility that such activation can be somehow avoided and this Whorfian effect can be reversed as well. Therefore, the Whorfian effect noticed by Deutscher should be additionally researched. Nevertheless, this Whorfian effect is already a good opportunity for these scientists, who reject reversible Whorfian effects. Properties of this Whorfian effect should also be useful in formulating the scientifically testable thesis of the Sapir-Whorf hypothesis.

5. Conclusion

In this MA thesis I realized three goals. Firstly, I defined properties of reversible Whorfian effects. Secondly, I showed that reversibility and its properties should be necessarily taken into account in research on whether or not empirical data can support the Sapir-Whorf hypothesis. Otherwise, reversibility can be easily mixed up with the reason to reject reversible Whorfian effects as empirical evidence for the Sapir-Whorf hypothesis without any addition justification. Thirdly, I showed that the Whorfian effect by Deutscher seems to be irreversible and should at least become an alternative for these researchers who reject reversible Whorfian effects as evidence for the Sapir-Whorf hypothesis. Summarizing all the defined properties of reversible and irreversible Whorfian effects, reversible Whorfian effects often includes nudge, can be reversed by change of background, detection in real life is unlikely and repetition in different experiments is also unlikely. Irreversible Whorfian effects never include nudge, are insensitive to the change of background, it's likely they can be repeated in different experiments and that they can appear in real life. Therefore, I would formulate reversible Whorfian effects as Whorfian effects, which can appear mostly in narrow laboratory situation and unlikely in real life. At the same time, I would define irreversible Whorfian effects as Whorfian effects, which can be repeatable in different laboratory situations, and are as likely to constantly appear in real life. All the reversible and irreversible Whorfian effects can be reversed in really special cases, e.g. in childhood or probably by study of a foreign language. Therefore, Whorfian effects can be demonstrated only by language users of appropriate level, e.g. by adults, native speakers, not having mental diseases etc. Irreversible Whorfian effects seem to be much stronger empirical evidence for the Sapir-Whorf hypothesis, than reversible Whorfian effects. At the same time, it's difficult to prove that irreversible Whorfian effects, like this irreversible Whorfian effect noticed by Deutscher, can't be somehow reversed in the future. Nevertheless, both definitions of reversible and irreversible Whorfian effects with their defined properties should be useful for philosophical research on formulating of the scientifically testable thesis of the Sapir-Whorf hypothesis.

Summary

Title: Reversible and irreversible Whorfian effects as empirical evidence for the Sapir-Whorf hypothesis

Pealkiri: Sapiri-Whorfi hüpoteesi pöörduvad ja pöördumatud efektid kui selle empiiriline tõestus

This thesis is aimed on research of two different types of empirical evidence for the Sapir-Whorf hypothesis: reversible and irreversible Whorfian effects. Definition of reversible Whorfian effects and their properties should help to identify reversible Whorfian effects easier in future research on the Sapir-Whorf hypothesis. I presented two opposite points of view on whether or not reversible Whorfian effects can support the Sapir-Whorf hypothesis to show that reversible Whorfian effect are actively disputable nowadays. Also, I showed that without a clear understanding of what reversible Whorfian effects are some researchers can mix up reversibility with the reason to reject Whorfian effect without any additional justification. This MA thesis should help to avoid this misunderstanding in future research. Moreover, I introduced a really good candidate for irreversible Whorfian effect and defined its properties. It's really difficult to prove that a certain Whorfian effect is totally irreversible, but my candidate at least should be a good opportunity for these researchers, who reject reversible Whorfian effects as empirical evidence for the Sapir-Whorf hypothesis. This MA thesis in general should help in philosophical research on formulating the scientifically testable thesis of the Sapir-Whorf hypothesis.

References

Ahearn, Laura M. (2011). *Living Language: An Introduction to Linguistic Anthropology*. John Wiley & Sons. ISBN 978-1-4443-4054-9.

Bennett, J. (1964/1989). *Rationality: An essay towards an analysis*. New York, Hackett.

Bennett, J. (1988). Thoughtful brutes. Proceedings and Addresses of the American Philosophical Association, *APA Eastern Division Presidential Address, in APA Proceedings 62*. Nevark, University of Delaware: 197–210.

Berlin, B, & Kay, P. (1969). *Basic color terms: their universality and evolution*. Berkeley and Los Angeles, California: University of California Press.

Boroditsky, L., L. Schmidt, and W. Phillips. (2003). Sex, Syntax and Semantics, *Language in Mind: Advances in the Studies of Language and Cognition*. Ed. Gentner and Goldin-Meadow: Cambridge, MIT Press.

Brown, P., and S. C. Levinson. (1993). "Uphill" and "downhill" in Tzeltal, *Journal of Linguistic Anthropology* 3: 46–74.

Carruthers, P. (2002). The Cognitive Functions of Language. *Behavioral and Brain Sciences 26*: 657–73. URL: https://www.researchgate.net/publication/227689525_Uphill_and_Downhill_in_Tzeltal

Davidson, D. (1975/1984). Thought and talk, *Inquiries into truth and interpretation*. Oxford, England, Oxford University Press: 155–170.

Davidson, D. (1982/2001). Rational animas, *Subjective, intersubjective, objective*. Oxford, England, Oxford University Press: 95–105.

Deutscher, G. (2010). *Through the Language Glass. Why the World Looks Different in Other Languages*: New York.

Fodor, J. (1975). *The Language of Thought*: Cambridge, Harvard University Press.

Gilbert, A., T. Regier, P. Kay, and R. Ivry. (2006). Whorf hypothesis is supported in the right visual field but not the left, *Proceedings of the National Academy of Sciences 103 (2)*. Washington: 489–94.

Gordon, P. (2004). Numerical Cognition Without Words, *Evidence From Amazonia*, *Science 306*: 496–499. URL: https://pdfs.semanticscholar.org/c6af/60ea50ac17fd8a879daca57fe333e999d816.pdf

Hauser, M. D. (2000). *Wild Minds: What Animals Really Think*: New York, Henry Holt.

Heine SJ. (2008) *Cultural Psychology*: New York, W. W. Norton.

Hermer-Vazquez L., E. Spelke, and A. Katsnelson. (1999). Sources of Flexibility in Human Cognition: Dual-Task Studies of Space and Language, *Cognitive Psychology 39*: 3–36, URL: https://www.researchgate.net/publication/12865560_Sources_of_Flexibility_in_Human_Cognition_Dual-Task_Studies_of_Space_and_Language

Imai, M., & Gentner, D. (1993). Linguistic relativity vs. universal ontology: Cross-linguistic studies of the object/substance distinction, *Proceedings of the Chicago Linguistic Society*, URL: http://groups.psych.northwestern.edu/gentner/newpdfpapers/ImaiGentner93.pdf

Imai, M., & Gentner, D. (1997). A crosslinguistic study of early word meaning: Universal ontology and linguistic influence, *Cognition 62*: 169–200, URL: https://www.sciencedirect.com/science/article/pii/S0010027796007846

Kay, P. and W. Kempton. (1984). What is the Sapir–Whorf Hypothesis?, *American Anthropologist 86*: 65–78, URL: https://www1.icsi.berkeley.edu/~kay/Kay&Kempton.1984.pdf

Kosslyn, S. M., W. L. Thompson, and G. Ganis. (2006). *The Case for Mental Imagery*: Oxford, Oxford University Press.

Lemer, C., S. Dehaene, E. Spelke and L. Cohen. (2003). Approximate Quantities and Exact Number Words: Dissociable Systems, *Neuropsychologia 41*, 1942–1958, URL: https://www.harvardlds.org/wp-content/uploads/2017/01/lemer2003-1.pdf

Levinson, S. C., S. Kita, D. B. M. Haun, and B. H. Rasch. (2002). Returning the tables: Language affects spatial reasoning, *Cognition 84*: 155–88, URL: https://www.sciencedirect.com/science/article/pii/S0010027702000458

Lecours, A. R., and Y. Joanette. (1980). Linguistic and Other Psychological Aspects of Paroxysmal Aphasia, *Brain and Language 10*: 1–23, URL: https://www.sciencedirect.com/science/article/pii/0093934X80900346

Li, P., and L. Gleitman. (2002). Turning the tables: Language and spatial reasoning, *Cognition 83*: 265–294, URL: https://pdfs.semanticscholar.org/12a6/61f7f6a1b35e333e30328a9ff4d268812e74.pdf

Lucy, J. (1992). *Grammatical categories and cognition: A case study of the linguistic relativity hypothesis*: Cambridge, England, Cambridge University Press.

Lucy, J., & Gaskins, S. (2001). Grammatical categories and the development of classification preferences: A comparative approach, *S. Levinson & M. Bowerman (Eds.). Language acquisition and conceptual development*. Cambridge, England, Cambridge University Press: 257–283.

Pederson, E., Danziger, E., Wilkins, D. G., Levinson, S. C., Kita, S., & Senft, G. (1998). Semantic typology and spatial conceptualization, *Language 74*: 557–589, URL: https://www.jstor.org/stable/417793?seq=1#page_scan_tab_contents

Pöhls, R. L. V. (2013). Testing the untestable? Guidelines for advancing empirical research in the area of Linguistic Relativity, *RIFL vol. 7 n. 3*: 98–108, URL: http://www.rifl.unical.it/index.php/rifl/article/view/178

Quine, W. V. O. (1960).*Word and object*: Cambridge, MIT Press.

Reines, Maria Francisca; Prinz, Jesse (2009). Reviving Whorf: The Return of Linguistic Relativity, *Philosophy Compass. 4 (6)*: 1022–1032, URL: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1747-9991.2009.00260.x

Schaller, S. (1991). *A Man Without Words*: New York, Summit Books.

Sera, M. D., Berge, C., & del Castillo Pintado, J. (1994). Grammatical and conceptual forces in the attribution of gender by English and Spanish speakers. *Cognitive Development 6*: 119–142, URL: https://www.sciencedirect.com/science/article/pii/0885201494900078

Tan, L. H., A. H. D. Chan, P. Kay, P. L. Khong, L. K. C. Yip, and K. K Luke. (2008). Language affects patterns of brain activation associated with perceptual decision, *Proceedings of the National Academy of Sciences 105 (10)*: 4004–4009, URL: https://www.researchgate.net/publication/5534980_Language_affects_patterns_of_brain_activation_associated_with_perceptual_decision

van Troyer, G. (1994). Linguistic Determinism and Mutability: The Sapir-Whorf "Hypothesis" and Intercultural communication, *JALT Journal, 16(2)*: 163–178, URL: https://www.jalt-publications.org/jj/articles/2761-linguistic-determinism-and-mutability-sapir-whorf-hypothesis-and-intercultural-comm

Weiskopf, D. A., Adams, F. (2015). *An introduction to the philosophy of psychology 316*: Cambridge, Cambridge university press. ISBN 978-0-521-74020-3

Winawer, J., N. Witthoft, M. C. Frank, L. Wu, A. R. Wade, and L. Boroditsky. (2007). Russian blues reveal effects of language on color discrimination, *Proceedings of the National Academy of Sciences 104 (19)*: 7780–7785, URL: https://www.researchgate.net/publication/6358875_The_Russian_Blues_Reveal_Effects_of_Language_on_Color_Discrimination

Кронгауз, М. А. 2001 — Семантика. М.

Кронгауз, М. А. 2011 — Язык. Мышление. Коммуникации. *Открытая лекция в МПУ*. URL: https://www.youtube.com/watch?v=vGfIIVDyiGQ

Кронгауз, М. А. 2012 — Гипотеза лингвистической относительности. *ПостНаука*. URL: https://postnauka.ru/video/6759

Non-exclusive licence to reproduce thesis and make thesis public

I, Maksim Grigorev,
(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

REVERSIBLE AND IRREVERSIBLE WHORFIAN EFFECTS AS EMPIRICAL EVIDENCE FOR THE SAPIR-WHORF HYPOTHESIS,
(title of thesis)

supervised by Alexander Davies and Bruno Mölder,
(supervisor's name)

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 14.05.2018