MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"Unconscious Inference: Understanding Semantics Through Colour Concept Processing in Verbal and Visual Modalities"

verfasst von / submitted by

Vadim Kulikov PhD

angestrebter akademscher Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2016 / Vienna 2016

Studienkennzahl It. Studienblatt / Degree programme code as it appears on the student record sheet:	A 066 013
Studienrichtung It. Studienblatt / Degree programme as it appears on the student record sheet:	Masterstudium Joint Degree Programme MEi:CogSci Cognitive Science
Supervisor:	Prof. Dr. Ulrich Ansorge

Памяти моего дедушки Николая Семёновича,
который первым открыл мне мир науки.

Abstract

This thesis explores various aspects of perception – what is it, how does it work and how is it related to meaning and subjective experience. Perception is taken to involve the perceiver, the outside world and the way different cognitive and sensory modalities interact to produce an experience. In particular experience and perception are not only determined by the external world, but also by the perceiver's internal world. A theory of cognitive semantics called *framework theory* is briefly presented. It attempts to explain the emergence of meaning and content as a result of such interactions; this may be considered as one of the contributions of the thesis.

As the main tools to further understand the mechanisms of perception, this thesis employs the Stroop and priming tasks. Both tasks produce phenomena in which the interference between different cognitive and perceptual processes can be detected and used to test theories and models of brain function. It appears that the Stroop effect can be in particular useful in testing the framework theory. A computational Hebbian neural network model that has its basis in FT was developed to explain a variety of experimental observations in the Stroop effect and its variations. This is the second contribution of this thesis.

The third contribution is motivated by research in categorical perception, the phenomenon in which experience is influenced by linguistic and non-linguistic concepts. A Stroop-like priming study was conducted to see if verbal processing interferes with perception more when stimuli are presented in the right visual field as opposed to the left visual field. Since the right visual field projects to the left cerebral hemisphere and vice versa, we expected to obtain more interference in the right visual field than in the left one, because the left hemisphere is generally linguistically more dominant. This turned out to be the case for a priming setup where both prime and target are non-verbal colour stimuli. For the setup where the primes were verbal, the results are still somewhat unclear, but there seems to be a stronger effect in the combination of right eye with right visual field than in any other combination of eyes with visual fields.

This is an interdisciplinary work where philosophy of mind, computational cognitive neuroscience and cognitive psychology collide.

Zusammenfassung

Diese Arbeit untersucht verschiedene Aspekte von Wahrnehmung, was sie ist, wie sie funktioniert und wie sie mit der Semantik und den subjektiven Erfahrungen in Beziehung steht.

Der Grundidee nach ist Wahrnehmung das Resultat einer komplexen Kombination aus den Informationen der Umwelt, den Anlagen der Wahrnehmenden und den Interaktionen verschiedener kognitiver und sensorischer Modalitäten, die eine Wahrnehmung produzieren. Eine Theorie der kognitiven Semantik, „framework theory“, wird kurz präsentiert, welche das Aufkommen von Bedeutung und Inhalt als das Resultat solcher Interaktionen beschreibt. Dies kann als erster Beitrag eines neuen Denkansatzes interpretiert werden.

In dieser Arbeit kommen die Stroop- und Priming-Effekte zur Anwendung, um die Mechanismen der Wahrnehmung besser zu verstehen. Dies sind beides Phänomene, bei denen die Interferenz der Wahrnehmungs- und Kognitionsprozesse gemessen und als Testwerkzeug von Theorien und Modellen der Gehirnfunktion verwendet werden kann.

Insbesondere der Stroop-Effekt hat sich als geeignet herausgestellt, um die „framework theory“gewissermaßen zu testen. Ein auf dieser Theorie basierendes künstliches Hebbsches neuronales Netzwerkmodell wurde entwickelt, um einige der experimentellen Beobachtungen in den Stroop-Aufgaben zu erklären. Dies ist der zweite Denkansatz in dieser Arbeit.

Der dritte Denkansatz ist durch die Forschung in kategorischer Wahrnehmung motiviert. Kategorische Wahrnehmung beschreibt wie linguistische und andere Konzepte Erfahrung beeinflussen. Eine Stroop-ähnliche Primingstudie wurde durchgeführt, um festzustellen, ob verbale Verarbeitungsprozesse die Wahrnehmung mehr beeinflussen, wenn sie im rechten Wahrnehmungsfeld auftauchen, als sie dies im linken Wahrnehmungsfeld tun. Da das rechte visuelle Wahrnehmungsfeld über die linke Hemisphäre verarbeitet wird und umgekehrt, wurde im rechten Blickfeld mehr Interferenz erwartet. Dies war auch der Fall in einem Priming-Setup, in dem sowohl Hinweisreiz als auch Zielreiz nonverbale Farbstimuli waren. In einem Setup mit verbalen Hinweisreizen sind die Resultate noch immer nicht eindeutig, dennoch scheint in der Kombination von rechtem Auge mit rechtem Blickfeld ein stärkerer Effekt aufzutreten, als in irgendeiner anderen Kombination von Auge und Blickfeld.

Dies ist eine interdisziplinäre Arbeit, in der die Philosophie des Geistes, die informationsverarbeitende kognitive Neurowissenschaft und die kognitive Psychologie kollidieren.

Contents

Acknowledgements

I wish to express my gratitude to my supervisor Prof. Ulrich Ansorge for teaching me what cognitive psychology is all about and without whom this thesis wouldn't be possible. I am grateful to all the other teachers of the MEi:CogSci program (and beyond) whose lectures I visited in Vienna and in Bratislava. Especially I thank Elisabeth Zimmermann for her patience and encouragement as an organiser of the program. Last but not least, thanks to my dear friends in Vienna all of whom have extraodinary minds, especially Michael Schlattl with whom we inspired each other, and Kami Reiter with whom we loved each other.

Copyrights and Software

Only freely available software was used:

- LaTeX for the document preparation,
- GNU Emacs for typesetting,
- Python with Pygame for programming the interface of the experiment and for data analysis,
- R for the statistical analysis of the data,
- GIMP and METAPOST for image editing and creation.

The code for the experiment design (Section 9) and the data are available upon request by contacting me at vadim.kulikov@iki.fi.

The following figures were produced by the author and are not subject to copyright:

- Figures 4 and 5 on pages 26 and 31 were created using METAPOST.
- Figures 7 and 8 on pages 40 and 41 are scanned drawings edited in GIMP,
- Figure 11 on page 47 consists of screenshots from the experiment's GUI implemented in Python and Pygame.

The image on page 12 is published by Nature Publishing Group and is licensed out under a Creative Commons license and allows re-use and sharing for non-profit research purpose.

All other reproduced graphic images are published in the United States of America and are subject to "Fair use" according to the United States copyright law ("Nonprofit, educational, scholarly or research use").

All reproduced material is accompanied with the indication of its source or copyright owner.

> *It is quite wrong to try founding a theory on observable magnitudes alone. In reality the very opposite happens. It is the theory which decides what we can observe.*
>
> A. Einstein[1]

1 Introduction

The main theme of this thesis is to explore the various ways in which our perception of the environment is shaped by what there is *in us* prior to the experience. Naturally, the anatomy and biological architecture of sensory organs play a crucial role in what and how we perceive, but that is not all. Beliefs, attitudes and expectations, prior and simultaneous experiences, all contribute to perception. Section 2 briefly explores a range of scientific findings that demonstrate how this was possible to know and what is understood about it, in terms of the philosophical and phenomenological perspectives, all the way to psychological and even statistical contributions. The section is concluded with a generalised theoretical account (framework theory) which argues that the idea of perceptual inference can be applied more granularly to individual senses and perceptual modalities (rather that individual agents) and that semantics and meaning do in fact emerge from this type of interaction between these modalities. Taken to extreme, this theory claims that not only do prior predispositions *influence* perception, but that they are, in fact, all there is to it and the sense of meaning emerges from their interactions; for one needs a brain to be able to perceive, but one cannot have a brain without having *some* predispositions.

Then the thesis moves onto the Stroop effect which is a very specific example of this phenomenon. The Stroop task allows the testing of very precise predictions about the nature of perception and semantic processing. In order to demonstrate this, I present an example of a typical Stroop task below. As you can see there is a sequence of words printed in coloured ink. Your tasks are to (1) read all the words normally out loud and (2) name the ink colour of each of the words:

blue green red blue red

If the effect worked with you, you notice that the latter exercise is more challenging and takes more time to complete. If you named the ink colours from left to right, you might have even experienced some slow-down effect at the last word where the colour is congruent with the word, because you already learned to "ignore" the conflicting information provided by the words earlier in

[1] In a discussion with Heisenberg about the nature of quantum mechanics. Quoted from Kumar (2008, Ch. 10). Originally from Heisenberg (1971).

the sequence. At the end of the section, a computational model of the Stroop effect is presented. This model is partially inspired by the framework theory and I will give both a theoretical analysis of the model and some data from the simulations to support its case.

The topic following will be on categorical perception which is another concrete example of how prior knowledge, namely linguistic and pre-linguistic categorisation, influences perception. The supervisor of this work, Prof. Ulrich Ansorge, and I are interested in exploring how much language is involved the processing of colours and in their categorisation. The Stroop effect already hints that there is a non-trivial connection. In order to study the involvement of language, we are interested in the lateral differences in reaction times to colours which we present either in the left or the right visual field. The verbal involvement hypothesis predicts that there is a difference in processing, because the left visual field projects to the right cerebral hemisphere which is not, in most humans, strongly involved in language processing, while for the right visual field the situation is reversed and the processing is initiated in the left, verbally dominant hemisphere. The results are not yet conclusive; our current hypothesis being that for a range of semantically different colours (e.g. red, blue and green) there is a lateral difference in processing, but for a range of semantically similar (e.g. different shades of blue) there isn't, which would mean that language indeed is part of categorical processing.

2 Experience and Perception

2.1 Inaccessibility of Subjective Experience

We all presumably possess *qualia*, the qualitative phenomenological units of experience. All sighted people should know what the blue colour looks like to them, how it appears to them and what the quality of the experience is i.e. what is their quale of the blue colour.

Nagel (1974) argues that the phenomenological experience of a living being is only accessible to that being itself and no-one else can have a complete understanding of it. He starts by arguing that we, as humans, cannot understand the experiential world of a bat. His choice of an animal is not coincidental. Bats are phylogenetically very close to us, humans; not only are they mammals, they are more closely related to humans than rodents (Novacek, 1992). This makes it difficult to doubt that bats do have at least *some* sort of phenomenal experience. But, alas, as Nagel puts it "Even without the benefit of philosophical reflection, anyone who has spent some time in an enclosed space with an excited bat knows what it is to encounter a fundamentally alien form of life." Most microbats (the suborder Microchiroptera or according to a more recent classification Yangochiroptera) navigate three-dimensionally using ultrasound echolocation. They do it successfully in flocks not getting confused by each others' sounds and the precision and quality are sufficient to catch insects on the fly and to distinguishing water bodies from other surfaces (Greif & Siemers, 2010). Recently it has been found that at least some bats have navigation cells in their hippocampus which code for three-dimensional information (Finkelstein et al., 2014). Nagel argues that in order to understand what it is like to be a bat, it is not sufficient to imagine that one has "webbing on one's arms, which enables one to fly around at dusk and dawn catching insects in one's mouth; that one has very poor vision, and perceives the surrounding world by a system of reflected high-frequency sound signals; and that one spends the day hanging upside down by one's feet in an attic." Nagel points out in a footnote that the English expression "what it is *like*" is already misleading, because "[i]t does not mean 'what (in our experience) it resembles' but rather 'how it is for the subject himself'" (Nagel, 1974). For example the German counterpart "Wie ist es, eine Fledermaus zu sein?" is more accurate.

Nagel concludes that phenomenal experiences are *subjective* while reductionist science tries only to talk about things in the world in an *objective* way which depends as little as possible on the observer. This qualitative gap, he argues, makes phenomenology unexplainable by current reductionist approaches.

Nagel concludes the paper with a "speculative proposal" that we should try to find an objective methodology to communicate and describe the phenomenal states which does not depend on imagination or empathy. Nagel calls it "objective phenomenology". It would amount to developing a new language or conceptual system capable to express subjective experiences so that eventually we could, using this language, describe what is it like to be a bat or describe to a blind person what it is like to see – at least better than is presently possible. Of course, he admits that this has limitations and for a reason. It may be that your experience of what we call blue is equivalent to my experience of what we call red and vice versa, but we do not notice this discrepancy, because we refer to the same external objects with the same (colour) names. It remains, in fact, a philosophical question how do we know that others have qualia at all; that's why I used the words "presumably" and "should" in the first two sentences of this section.

Investigating the questions *"How is it possible that two organisms have a different experience in the presence of the same stimulus?"* and *"Why can the experience of one organism be inaccessible to another?"* can illuminate the questions *"Why is there experience at all?"*, *"How does experience emerge?"* and *"What are the building blocks of experience?"*.

2.2 Perception as Inference

Last year the picture of Figure 1 went viral on the Internet. It is a photograph of a dress and it strongly divided opinions: some people claimed that the colours of the dress are white and gold and others purportedly experienced it

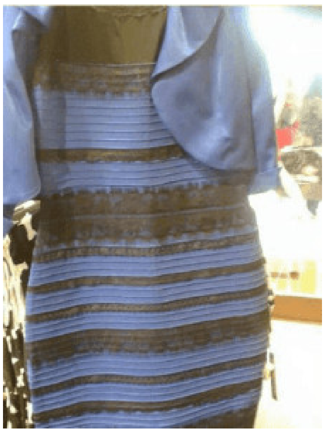


Figure 1: This picture of a dress[2] went viral on the Internet in 2015.

as blue and black. How is it possible that people perceive the same image in two (and maybe more) different ways?

The German scientist Hermann von Helmholtz described perception as *unconscious inference* (unbewusster Schluss). He hypothesised that what we think we perceive is not the same, and not isomorphic, the information that is actually flowing in from the senses (or what is "out there") and that the actual perception is the product of unconscious processes applied to this information. According to this view, if two people have different prior information in their brains (in this case probably whether the dress is in a shadow or in a bright sun light), then the same image can elicit different perceptions. This is, however, obviously quite rare. Or at least, in most cases it is hardly noticeable, for even if the same picture elicits different associations and feelings, still people from the same cultural background tend to describe in fairly similar terms what they see. But it does not mean that what we perceive is what there really is. Rare examples such as this picture of a dress and some visual illusions (see next page) reveal the fact that there is more to perception than mere "truth taking"[3].

The Bayesian Brain Hypothesis (BBH) is an attempt to formalise this view. The Bayes formula in statistics tells how to combine a prior probability distribution (corresponding to the information that is in the brain prior to the perception) with evidence (sensory data) to form a posterior probability distribution (the perception). There is a convincing body of evidence that (1) the brain is good in Bayesian estimates in natural situations (Ernst & Banks, 2002; Körding & Wolpert, 2006; Hoffrage & Gigerenzer, 1998) and that (2) people can have widely different prior distributions (Houlsby et al., 2013). One can conclude from (1) and (2) that the perceptions of different people can and

[2] Copyright Cecilia Bleasdale

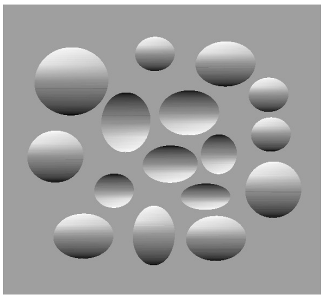
[3] *Perception* in German is *Wahrnehmung*, literally *truth taking*.

should be quite different.

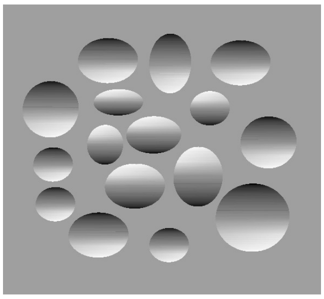
But it would be naïve to assume that the situation is so simple. Apart from prior information we also have particular embodiment. It is clear that a big role in shaping our experience is played by our sensory organs and their anatomy. Different sensory modalities and different sensory pathways may have different prior information and predispositions. If you place left hand into cold and right hand into warm water, then moving both of them into room temperature will result in different experience by different hands. In some sense the hands are differently primed.

Moreover the priors of different modalities can be influenced by each other: the expectations, and therefore prior information, of the auditory system can be influenced by the visual system which can result in hearing sounds differently depending on the visual scene (McGurk & MacDonald, 1976). Similarly, higher level cognitive mechanisms such as the knowledge of the present situation can influence perception in other modalities, for example the perception of pain is sensitive to higher mental processes (e.g. Wiech, Ploner, & Tracey, 2008; Garland, 2012). Even the different visual pathways, ventral and dorsal, are known to engage in independent processing of the visual sense data (Eysenck & Keane, 2000, p. 48), and so they may also have distinguishable priors. Finally, research on multisensory integration suggests that the experience of the world relies on statistically optimal integration of information from different senses (Ernst & Banks, 2002).

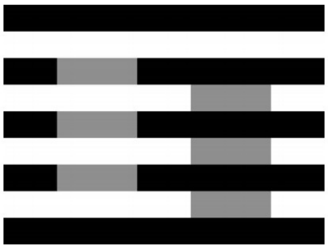
The prior information can be either innate, learned, or acquired shortly (in a matter of seconds or minutes) before the exposure to the stimulus. In the following picture people usually see concave bubbles surrounded by convex bubbles (Sun & Perona, 2008):



This image, however, is consistent with the opposite interpretation in which the reader can be convinced by turning this page upside down; the latter is not necessary, however, because the image is reproduced upside down right here:



If the illusion works for you, dear reader, then you see convex bubbles surrounded by concave bubbles in the latter, rotated picture and vice versa in the first picture. It is still under debate how much of this phenomenon is innate and how much is learned from experience (Sun & Perona, 2008), but it is certainly a combination of these two. The White's illusion in which the grey rectangles on the left look brighter than the ones on the right is known to be even more persistent and hence more likely to be innate (Adelson, 2000):



The grey rectangles have the same shade of grey. On the other hand, perceptual priors which are most certainly due to experience include the priors for face-perception which were studied by Houlsby et al. (2013). Priming effects where the perception of stimuli is influenced by the stimuli immediately preceding them are looked at more closely in Section 4, and categorical perception where learned (verbal) concepts influence perception is discussed in Section 5.

The pre-reflective beliefs about the way the world is, and the way one is immersed in it, are known to shape the perception in fundamental ways. For example the feeling and attitude that is triggered upon the presence of another person is such. This mechanism is so deeply integrated into our everyday experience that it is difficult to acknowledge its existence unless it is malfunctioning. The (usually terrifying) experience of such a malfunction is often described as seeing other people as “robots or puppets” (Sechehaye, 1970, p. 29). Phenomenologists refer to the totality of such mechanisms and the result of their function as *being-in-the-world* (Wheeler, 2013).

Thus, we have identified the following pre-perceptual entities that shape and influence the way sensory data is transformed into a percept or perceptual experience, possibly including qualia:

- Embodiment (physiology of sensory organs),
- Innate and learned high-level priors (allowing for (unconscious) inferences such as the concavity of a bubble),
- Pre-reflective attitudes and being-in-the-world,
- Verbal and symbolic categories (like in categorical perception),
- Multisensory integration,
- Priming effects (current perception influenced by the immediately preceding one).

Understanding these mechanisms and the way they shape perception can take us closer to Nagel's speculative proposal of describing phenomenal experiences. For example, if we know the prior distribution of faces of persons A and B (Houlsby et al. (2013) demonstrated that it is already, at least to some extent, possible), we could generate a mapping which would transforms a given face to look from the point of view of A the way it would normally look from the point of view of B.

2.3 Semantics and Frameworks

Rather than asking "Where experience comes from?" one can ask "Where meaning comes from?". I believe that answering the latter question is easier, but is also necessary before tackling the former question. Why are perceptions meaningful? How are meanings of symbols grounded?

Can there be pure perception? Perception that is not influenced by any prior predispositions? No. As argued in Introduction, one needs a brain in order to perceive and this brain must be have *some* architecture which implies some predisposition. But how then is it possible that we see a cat, then it is absolutely clear to us that it is a cat and the concept of cat seems to be quite pure in our minds. As if there was some sort of "pure" perception of a cat, a semantic representation of it, which is independent on the exact colour and size of, or even associations with, the currently perceived cat.

A classical account, or at least a general way of speaking, in cognitive psychology textbooks and research papers is that when one sees the word "cat", the semantics of cat is being evoked in one's mind. The word is said to be "semantically processed". By this, it is understood that some centralised concept of a cat is activated which is then responsible for the further cognitive flow: images and sounds of cats, memories related to cats, narratives and knowledge about cats and all the "semantic" priming effects known to follow the exposure to "catness" (cf. Section 4). But what if images, words, sounds, memories and priming effects is all there is, and the associations are between those modalities and not spreading from any centralised "cat concept"? What if the semantics of "cat" emerges from the combination of congruent and coherent patterns in different modalities? Cognitive scientists will notice that what I am doing is close to grounding symbols in one modality in symbols in another modality. Many philosophers have argued that it is impossible, for example Harnad (1990) calls it the "Chinese-Chinese dictionary" problem where someone who has no clue about Chinese is trying to learn it from such a dictionary. This problem, however, is arguably unavoidable at least on some level. The proposed "solutions" often involve grounding in the sensory-motor domain (Kiefer & Barsalou, 2013) which is essentially just changing the problem to a "Chinese-Malaysian dictionary" problem for someone who doesn't know neither Malaysian nor Chinese. Sensory-motor data consists of sequences of neural firings, which "Malaysian" in this case stands for. To appreciate that the sensory-motor domain is (a priori) equally meaningless as any other collection of symbols, think of a programmer who is developing a software for an embodied robot. She has to deal with the sensory-motor domain of the robot's perceptual mechanisms in the form of zeros-and-ones (or other strings of symbols) which are exactly the same kind of symbols that deal with the robots higher cognitive abilities. Unfortunately more detailed argumentation falls out of the scope of this thesis.

Above I used the word "modality" to refer to different modes of information processing. This, however, is a too narrow word, because, as seen above, I also include abstract knowledge and narratives about cats into the range of various ways to approach the concept. Originally I called them *frameworks* (Kulikov, 2015). From Section 2.2 we see that any coherent perception is a combination of embodied and cognitive information processing in which many frameworks are involved. Maybe this term is better understood through a different example. Consider the Milky Way. One can go into a dark place and look at the clear night sky and see the pale lane traversing the skies. This is one framework of looking at it. One can also go to Wikipedia and read about the Milky Way as being a flat disc with tentacles; or one can open an astrophysics textbook and try to understand quite complicated mathematical formalisms explaining (or failing to explain) some observed behaviours of the galaxy. These are three quite different ways of engaging with the same concept, but note that it is an *essential part* of the concept that all these three ways are available. The attempt is to explain how meaning can emerge when originally only meaningless (neural firing) data is available to the organism. If the organism had only one sensory modality (in extreme case, say, only one sensory neuron), no meaning can arise (unless there is an innate or inbuilt interpretation). But as soon as there are many sensory modalities, the information flowing from them can be compared and separate frameworks – ways of looking at the world – can emerge. Then it is possible to make truth judgements: if what I see through this lens matches all the other available lenses, then it is likely to be correct.

To illustrate this, let us look at another example: what would be sufficient to convince me that there is a living pink elephant floating in the air in my office in Vienna? If someone would try to convince me of this, I would first doubt their sincerity and then question their mental health – sooner than I would believe them. If I had seen the elephant floating, I would still rather question my own mental health than believe my eyes. But what if I could also touch it? And hear it? And what if, additionally to all this, I had a perfectly sound explanation of why the elephant is indeed floating in the air in my office? What if I had a narrative of who placed it there, how and why? If all this was the case I would actually believe it! I know this, because this is exactly the reason I believe right now that there is a table in front of me as I am writing this sentence. So the question of truth, according to this view, also boils down to coherence between different frameworks, or modalities, or one could also call them *compartments*.

Note also that framework theory is a kind of generalisation of theories attempting to explain symbol grounding in sensorimotor terms (Kiefer & Barsalou, 2013). Framework theory says that symbols in each framework are grounded in the other frameworks, some of which are the sensorimotor modalities. This generalisation step is in my view necessary to account for abstract concepts such as the infinite dimensional Hilbert space and to completely avoid semantic commitment (Cubek, Ertel, & Palm, 2015; Taddeo & Floridi, 2005). Further details fall unfortunately outside of the scope of this thesis.

Framework theory might seem to be vaguely posed, and it necessarily is so, because it needs to allow for generality. A precise model of a cognitive phenomenon (the Stroop effect) inspired by this view is described in Section 3.6.1. There, a phenomenon called "semantic interference" (alongside with other phenomena) is explained without a reference to a separate "semantic modality" as is often done in the literature, e.g. van Veen and Carter (2005) even identified a brain area responsible for it.

One can think of framework theory as a generalisation where each perceptual modality, as the ones identified in Section 2.2, acts like an individual perceiving information coming from other such modalities and giving back the result of its own processing. Each of them can always verify their "own beliefs" by comparing them to the "beliefs of others" and each of them has its own prior information and experience. Some of them are the result of coupling of more primitive ones. Perhaps the most primitive ones are the individual neurons and the least primitive ones are the individual brains, or groups of people, not excluding both bottom-up and top-down influence of concepts arising at different levels. The modalities of most interest further in this thesis are in the middle: e.g. sensory and verbal processing modalities.

3 Stroop Effect

Exactly one hundred years before I was born, in January 1886, the journal *Mind* published a paper entitled *The Time it Takes to See and Name Objects* by James M. Cattell, a student of W. Wundt. Cattell (1886) observes that the time needed to name a colour is about twice as long as the time needed to read a word. He suggests that this is because reading a word is "automatic", because the "association ... has taken place so often" but naming a colour requires a "voluntary effort to choose a name".

This spanned a wide interest in research of colour naming and word reading tasks as a way to study automaticity. Most of this research was focused on the differential-practice hypothesis which asserts, in line with Cattell, that the observed effect is due to much more practice in reading than in colour naming. Many studies were conducted during the following years with a lot of data which was interpreted either to support or to refute (and sometimes both!) the differential-practice hypothesis. Peterson, Lanier, and Walker (1925) presented an alternative hypothesis explaining longer reaction times for colour naming: that there are many different responses conditioned on a single colour, but only one response conditioned on a single word. MacLeod (1991) reviews a history of this.

One month short of half a century after Cattell's paper, in December 1935, John R. Stroop came across the idea that colour and word stimuli can be combined by printing colour names in coloured ink. For example the word "blue" in green ink colour. According to Google Scholar the resulting paper (Stroop, 1935) has been cited 13021 times by the time I am typing these words. In this study Stroop conducted three experiments, the first two of which were replicated in particular by MacLeod (1991) who obtained very similar results as Stroop. In the first experiment the participants were presented with words written in incongruent colours and were instructed to read the words out loud. The time needed to read 100 words was measured. The second experiment was identical except that the participants were instructed to name the colour of the ink in which the word was written. As a control condition in the first experiment words were written in black ink and in the second experiment coloured squares instead of words were used. Stroop observed that there was no significant interference effect in the first experiment, but there was a significant one in the second. A range of variations in the methodology in later articles (printing stimulus vertically rather than horizontally, using other shapes than squares in the control condition etc.; as summarised by MacLeod (1991, p. 166)) confirmed the robustness of the effect which persisted over all these modifications. Stroop himself proposed a theory of response competition according to which the brain generates responses both for the observed word and colour and the interference effect is due to competing responses from which one has to be picked (see also below). But this couldn't be the only explanation due to the asymmetry that Stroop also observed (significant interference only in the second experiment). Hence, he admitted that his findings are also consistent with the differential-practice hypothesis and the hypothesis of Peterson et al. (1925) about the multiple responses conditioned on colour but only one to text percepets.

Trying to find evidence in support of the differential-practice hypothesis, Stroop conducted another experiment where the participants practised over several days to name the colour of the ink. After the practice period, the incongruency effect in ink colour naming task was reduced and incongruency in the word reading task increased. This is currently called "the reverse Stroop effect", see also Section 3.3.

Later the idea of response competition of Stroop was broadened to semantic and task conflict theories and many theoretical models and explanations of various aspects of the effect have been introduced. Below the most prevalent of them are reviewed and in the end (Section 3.6.1) a Hebbian NN model is introduced which accounts for a wide range of data.

3.1 Different Types of Conflict

The response interference hypothesised by Stroop was later supported by experiments where the incongruent word either belonged to the set of possible responses or did not. The incongruence effect was weaker for those words which did not belong to the response set (Klein, 1964). Whereas in this model, the subject is supposed to generate two contradicting responses to the same stimulus, in semantic interference model it is assumed that the meaning of the written word is conflicting with the observed colour already before any response is being generated and the very generation of the response is being delayed. The task interference model, on the other hand, states that the perception of the written word elicits a task to read a word instead of the relevant task of naming the colour. There is a neurophysiological difference between the semantic and response interference as well as evidence that both indeed contribute to the Stroop effect (van Veen & Carter, 2005).

Using the paradigm where the subject has to respond to two different colours by pressing the same key, say a key on the left corresponds to yellow and blue and a key on the right to green and red, van Veen and Carter (2005) studied two types of interference. Semantic interference (SI) occurs between the written word and the ink colour when both would require the same response (for example "yellow" in blue ink). Response interference (RI) on top of semantic one occurs when the word denotes a colour which would require the other response (e.g. "red" written in green ink). In the third, congruent (CO) condition, the word denotes the same colour as the ink. In the fMRI image one can now isolate the brain areas in which the activation is stronger in the SI trials compared to the CO trials and where the activation is stronger in RI trials compared to the SI trials. The former is assumed to be responsible for the resolution of a semantic conflict and the second one for the resolution of the a response conflict. The authors concluded that there are different (in fact, disjoint) brain regions (in anterior cingulate, prefrontal, and parietal cortices respectively) responsible for the resolution of these two types of conflicts.

By task interference it is understood that when a word is presented in a visual field, the early processing stages automatically "think" that they have to perform the task of reading it thereby "forgetting" about the relevant task of recognising the colour, or alternatively that the time is simply spent on the decision on which task to perform. This hypothesis is supported by studies which show that in some situations even congruent stimuli produce more "interference" (longer reaction times) than neutral stimuli (Stirling, 1979).

3.2 Automatic and Voluntary Actions

The original account of Cattell (1886) on why word reading is faster than colour naming was that word reading is "automatic" and colour-naming a voluntary action. With this assumption it is relatively easy to also explain the Stroop-effect: if word-reading is automatic, it is faster and cannot be prevented from occurring even if it is unwanted. In this case, the semantic information of the colour is computed from the word already before it is computed from the ink colour resulting in the interference phenomenon.

This account has been challenged by many researches. For example it has been shown that the interference in ink naming task diminishes, or even disappears, if instead of the whole word, only one letter is coloured (Besner, Stolz, & Boutilier, 1997). However, they did not make a control experiment where one letter in a word looks different than the rest – e.g. one letter is bold or in another font. If reading a word becomes slower or less automatic in such a condition, then their claim that

> [t]his outcome flies in the face of any automaticity account in which specified processes cannot be prevented from being set in motion (Besner et al., 1997)

cannot be justified. To see this, imagine that the word reading is slowed down by changing the appearance of one letter. This means that in this situation word reading is no longer an automatic process and is not set in motion involuntarily to the same extent as when all the letters look the same. Thus, it would not be surprising that the word interferes *less* when one letter has a different colour.

3.3 Automaticity as Continuum and Reverse Effects

Automaticity seems to be often understood in different ways, and some authors misunderstand each other. For example the way in which Cattell (1886) expressed automaticity in this context in his early paper was misunderstood in the following passage:

> ...the automaticity account, which was rooted in Cattell's (1886) work.... Here the basic idea is that processing of one dimension requires much more attention than does processing of the other dimension. (MacLeod, 1991)

Whereas Cattell only wrote that it is the association of written words to a spoken words that is more automatic than the association of the ink colours to the spoken words. To say that the processing of colours *within* the colour domain is not automatic, voluntary or attention requiring is to say that when you open your eyes, the image you see is in black and white unless you specifically pay attention to the colours, which is obviously false. To be fair to MacLeod, however, he continues with:

> Thus, naming the ink color draws more heavily on attentional resources than does reading the irrelevant word. (MacLeod, 1991)

This in my view is already a different claim, which is more likely to be true. This distinction will be obvious in the NN model offered in Section 3.6.1. In particular it is different to say that processing of a stimulus is automatic or that a translation of stimulus in one modality to another modality (e.g. text to spoken word) is automatic.

MacLeod and Dunbar (1988) argued for the continuity of automaticity. They conducted an experiment where the subjects had to learn to attach names to newly learned shapes and these names were, in fact, colour words, see Figure 2. After that the shapes were presented in colour ink which was either congruent or incongruent with the name of the shape. The subjects had to either name the shape or name the ink colour after 1, 5 and 20 sessions of practicing the names of the new shapes. The results showed that the more they trained the more the shape interfered with colour naming. In the reverse task of naming the shape, the interference got weaker with practice.

Justified by the results of MacLeod and Dunbar (1988), Cohen, Dunbar, and McClelland (1990) introduced a continuous concept *strength* of a process which is supposed to replace the binary concept of automaticity by a continuous concept so that some process can be stronger than another instead of just being either automatic or non-automatic. They argue that if automaticity is such a binary concept, then from Stroop's classical findings one should conclude that colour naming is non-automatic, but from the findings of MacLeod and Dunbar (1988) one should conclude that it *is* automatic, because after one session of studying the new shapes, ink colour has the same role to the shapes as the words have to the ink colour in the standard Stroop task.

A completely reverse effect was obtained in an experiment conducted by Durgin (2000) where instead of naming the ink colour, the subjects had to

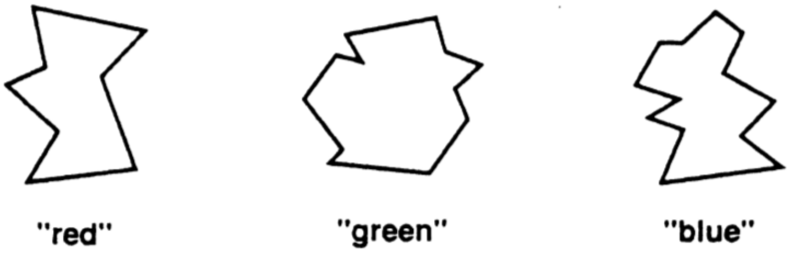


Figure 2: Shapes used by MacLeod and Dunbar (1988).

point to a square coloured with the same ink as the word's letters. Durgin rightfully points out that

> Accounts of Stroop interference that depend on the purported automaticity of verbal processing of text, ..., are difficult to adapt to the present results. (Durgin, 2000)

As a theory which is compatible with his findings, he recovers the "translational model" introduced by Virzi and Egeth (1985). These findings challenge virtually all previous theories (at least previously mentioned in this text) in that they show that the response modality plays a crucial role in the results and as mentioned above it is because *translation* and not *processing* (what does it even mean?) is the core notion here. None of the models described so far take the output modality into account and that's why are unable to explain the reverse effect. If the subjects are required to report the answers verbally, as in the original study by Stroop, then the classical effect is obtained: interference in ink-recognition of incongruent colour words, but little or no interference in reading the words even if they are written in incongruent ink colour. But if the answer has to be reported by, say, pointing to a colour patch, then the effect is reversed: there is an interference effect when the colour word has to be recognised, but little or no interference when the ink colour needs to be recognised. Thus, any account which doesn't explicitly include the response modality is necessarily incomplete.

3.4 Translational Model

The translational model by Virzi and Egeth (1985) assumes that there are separate cognitive systems processing different types of stimuli. For example in the classical Stroop task the linguistic system and visual colour processing systems are assumed to be involved; denote these two systems by L and C for linguistic and colour respectively. The stimulus is processed by both: the word by L and the colour by C. The output is verbal, so it takes place in L. Therefore if the task is to read the word, then there is no need for translation from C to L, the output being in the same modality as the relevant input. If the ink colour needs to be detected, then there is a need for translation from C to L, because the relevant stimulus is in C, but the output needs to be produced in L. This accounts for the asymmetry in the Stroop interference, which is then reversed, if the output has to be done in C. Of course, we do not have a natural modality for "colour output", so Virzi and Egeth design an experiment where both outputs have a similar status. Subjects were required to perform a card sorting task. Stimulus printed on the cards was either a colour word printed in black, a string of coloured X's or a colour word printed in a congruent or incongruent ink colour. The task was to sort the cards either according to ink colour or the word meaning into two bins which were either labelled by a colour word in black ink or a colour patch. The results they obtained were symmetric: when the bins were labelled by colour word, then the task where incongruently coloured words needed to be sorted by ink colour was the only task with significantly increased RT's and when the bins were labelled with colour patches, then the task where incongruently coloured words needed to be sorted by word meaning was the only task with significantly increased RT's.

Virzi and Egeth performed three other experiments also confirming the translational hypothesis and included a brief literature review showing that this model can also account for previously obtained data.

The drawbacks in this model are the following. First, it does not explain the semantic nor task interference described above, or at least the difference between them (van Veen & Carter, 2005; Stirling, 1979). Further, it does not explain the interference effect in a situation where both stimulus modalities are non-verbal, but the output is verbal as in the study of MacLeod and Dunbar (1988).

3.5 Parallel Distributed Processing Model

Cohen et al. (1990) develop a model which can be thought of as an extension of the translational model, although they didn't explicitly state it in the paper (they don't cite Virzi and Egeth (1985)). It is based on *backpropagation*, an idea stemming from machine learning and computational cognitive neuroscience. They call it *parallel distributed processing model* (PDP) because the processes of colour and word recognition are assumed to occur in parallel in a system where information is distributed over multiple units (neurons). Their model is a layered feed-forward artificial neural network (ANN) with three layers. The output-layer contains a neuron for each possible colour. The input layer is divided into three groups of neurons I_{ink}, I_{word} and I_{task} and the hidden layer into the two groups H_{naming} and $H_{reading}$. The group I_{ink} contains a neuron for every possible ink-colour and I_{word} a neuron for all possible colour words. The group I_{task} contains two neurons which specify whether the

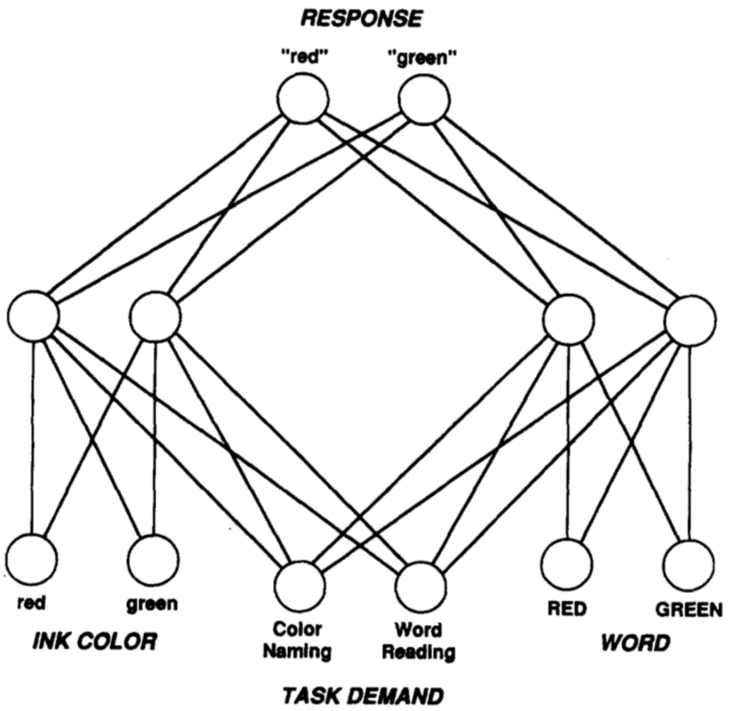


Figure 3: The network architecture of the PDP model of Cohen et al. (1990).

task is to name the ink colour or to read the word. The neurons in I_{ink} are only connected to neurons in H_{naming} and neurons in I_{word} only to neurons in $H_{reading}$. The neurons in I_{task} are connected to all hidden neurons and all hidden neurons are connected to all output neurons. The ANN is trained to “read” and to “name” colours using backpropagation algorithm: “reading” means that when the neuron in I_{task} corresponding to the reading task is activated together with the neuron corresponding to the word “green” in I_{word}, then the output neuron corresponding to green should also activate after the forward-propagation and so on, Figure 3.

Then, when confronted with the actual Stroop task, the input is administered from both sites, I_{ink} and I_{word} corresponding to the situation where the participant is presented with a word written in some ink colour. Then this input is forward-propagated through the ANN so many times as is needed for one of the output neurons to cross a given threshold. Once this happens, then the output is obtained. The number of needed iterations is supposed to model the reaction time.

They showed that if the network is trained a lot with the reading task, but only a little bit with the colour-naming task, then the standard Stroop effect and the related asymmetric pattern can be replicated.

This model replicates of course also the findings of MacLeod and Dunbar (1988) where different shapes were presented in colours, because this backpropagation model is blind to what are the actual stimuli, because it is a matter of interpretation of the neural activations which makes sense only from the "outside" point of view. The pathway which is trained the most is going to exhibit "more" automaticity or strength than the other. According to this view

> [t]he speed and accuracy with which a task is performed depends on the speed and accuracy with which information flows along the appropriate processing pathway. (Cohen et al., 1990)

3.6 Framework Theoretic Approach

Already Stroop (1935) noted the prevailing asymmetry of the task manifested in that reading colour words written in incongruent ink is not significantly slower than reading them in black ink, but naming colours is significantly slower if the ink is forming an incongruent colour word. An even more basic asymmetry was, in fact, the original finding by Cattell (1886) that simply naming colours takes twice as much time as reading words. As is seen from above sections, this asymmetry is sometimes interpreted as a difference in automaticity or comprehension. For example Melcher and Gruber (2009) write

> Thereby, Stroop-interference denotes the phenomenon that an irrelevant but incongruent word identity produces substantial crosstalk in the processing of the task-relevant colour. (Melcher & Gruber, 2009)

This quote seems to presuppose that the word somehow influences the actual (perhaps even visual) processing of the colour.

The translational model (Virzi & Egeth, 1985) and the PDP model (Cohen et al., 1990) described above are the closest in spirit to framework theory (Section 2.3), because it assumes the independent existence of various cognitive systems which interact with each other through "translation" and which otherwise act independently of each other. These models, however, are incomplete and do not explain all the available data; for example the PDP model does not explain the semantic interference as observed when the response to the semantically conflicting stimuli is the same (van Veen & Carter, 2005), and neither of the models seems to explain the fact that the presense of more congruent stimuli among incongruent ones increases the interference. In this section I would like to introduce a model which is inspired by all the models described above, combines them all together (in a sense) and which gives predictions of the magnitude of the interference effect in various situations; these predictions are of course subject to tests. This model can be thought of as an elaboration on the translational model of Virzi and Egeth or the PDP model of Cohen et al., but is also naturally seen as a consequence of framework theory.

3.6.1 Association Based Hebbian ANN Model

In each of the experimental paradigms and setups above we can isolate the following "frameworks": a visual text processing modality, a visual colour processing modality, a meta-cognitive modality controlling the attention (like the I_{task}-neurons in the PDP model) and an output modality. This is the minimal setup. The output modality can be for example the speech producing modality (which should be separated from the written text-processing modality) or it could be motor or visuo-motor modality if the output is in the form of a button press or pointing. Motivated by the "semantic interference" discussed before one could be tempted to add also the "semantic framework" where colours are represented perhaps amodally or semantically with links to all the other modalities, but this seems unnecessary to explain all the data.

In framework theory all these modalities would have their own concepts and there would be some (learned) coherence patterns between them and mechanisms to translate from one modality to another. In this case we can assume the simplest possible setup where each modality is represented by N neurons corresponding to the N different colours that are participating in the task to be modelled. There are connections between the neurons of different modalities whose strengths correspond to how strongly the concept in one modality is associated with the corresponding concept in the another. The strength of this connection represents the strength of the process of translation from one modality to another; this corresponds to the "strength" as defined by Cohen et al. (1990): the more automatic the connection is, the stronger it is. This notion suggests a continuous real valued range of possible values. In Hebbian learning this number will automatically be proportionate to the amount of simultaneous activation of the neurons, i.e. proportionate to training accounting for the differential-practice hypothesis which also supported by the evidence presented by Cohen et al. (1990) and MacLeod and Dunbar (1988).

Thus, the weights of the network are updated using a simple incremental Hebbian learning instead of the complicated backpropagation through a hidden layer which is biologically also less plausible (e.g. Bekolay, 2011) and does not clearly model the strength-phenomenon (only indirectly).

In the simplest possible situation, the Stroop's original experiment with colours and words as percepts and verbal output modality, the ANN would look like depicted on Figure 4 (next page). Notice that apart from the hidden layer neurons it is very similar to the one in the PDP model, Figure 3.

One of the differences to the PDP model is that there is also a connection between the colour processing and visual text processing modalities which is

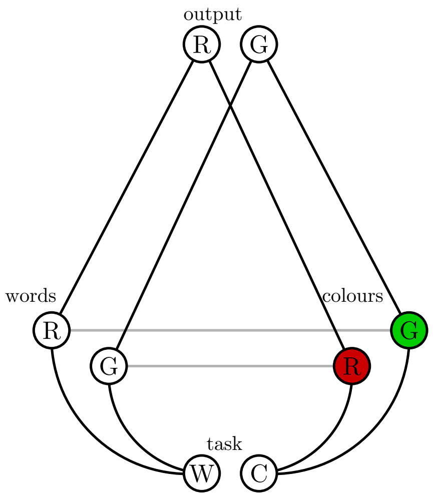


Figure 4: The architecture of the ANN to model the classical Stroop effect.

natural, since people do learn to associate also written words with colour, not only spoken words. This model is supposed to replicate (and indeed it does, see Results section below) the standard Stroop effect in the same way and essentially for the same reasons as the PDP backpropagation model of Cohen et al.. Let us look at the model more precisely.

3.6.2 Technical Description

The model is a continuous time recurrent neural network (CTRNN). In the first part of the experiment the architecture is as depicted on Figure 4. In reality there are four colours in each modality, but in the picture only two are depicted for clarity. Two neurons are coloured just to indicate that they are representing the colour processing modality. Each neuron can be in a state from -5 to 5 and each connection can have a weight from -5 to 5. The output of each neuron equals $\sigma(s)$ where s is the state of the neuron and σ is the activation function

$$\sigma(x) = \frac{1}{1+e^{-x}}.$$

The dynamics of the CTRNN is governed by the following equation

$$\dot{s}_i = -\tau s_i + \sum_j w_{ji}\sigma(s_i) + \eta I_i \tag{1}$$

where τ and η are constants, $\dot{s}_i$ is the time derivative of the state of the i:th neuron, w_{ji} is the weight of the connection from neuron j to neuron i and I_i is the external input to the neuron. This is a classical approach to modelling e.g. minimal cognitive agents (e.g. Beer, 1997; Beer & Williams, 2014; Maniadakis & Tani, 2009) but is also similar to the model of Cohen et al. (1990). Cohen et al. don't have a continuous time model; instead they define the state of each neuron as a weighted average of net inputs over time which becomes equivalent (up to constants) with the CTRNN model when the time is (necessarily) discretised for the purpose of the implementation.

The difference to both models is that here I use neither backpropagation (like Cohen et al.) nor evolutionary learning (like Beer; Beer and Williams.) but simple Hebbian learning. In the training phase (but ideally not only in the training phase, but always; see next section for a theoretical analysis), additionally to Equation (1), we apply the following equation to update the weights:

$$\dot{w}_{ij} = \alpha\sigma(s_i)\sigma(s_j) - \delta \tag{2}$$

where α is the learning rate and δ is the discount factor. The intuitive idea is that when both neurons, s_i and s_j are very active, then the connection w_{ij} increases rapidly, but when they are both small, it stays unchanged, or even decreases, if the term $\alpha\sigma(s_i)\sigma(s_j)$ goes below δ.

When discretised to time units of length 1, these equations become:

$$s_i(t+1) = s_i(t) + (-s_i(t) + \sum_j w_{ji}\sigma(s_i(t)) + \eta I_i)\tau dt \tag{3}$$

and

$$\dot{w}_{ij}(t+1) = w_{ij}(t) + (\alpha\sigma(s_i)\sigma(s_j) - \delta)dt \tag{4}$$

In my model the parameters were

$$\begin{aligned} \alpha &= 0.8 \\ \delta &= 0.04 \\ \tau &= 0.5 \\ \eta &= 0.5 \\ \eta &= 4 \\ dt &= 0.001. \end{aligned}$$

The response was obtain in the same way as in the NN model of Cohen et al. (1990) except that I didn't use noise. In the beginning of a trial the output vector μ is set to $\bar{0}$. For example in the case when there are four output neurons we have $\mu = [0, 0, 0, 0]$. Then at each step it is updated as follows:

$$\mu_i = \beta(\sigma(s_i), \max_{j \neq i} \sigma(s_j)), \tag{5}$$

meaning that the i:th coordinate of μ becomes the difference between the activation of the i:th output neuron and the output neuron which is most active apart from the i:th one. In our case $\beta = 0.01$. If μ_i crosses the threshold $\theta = 1$, $\mu_i > \theta$, then the output is i.

One can approach the experiment in two ways. One way is to simply adjust the weights between the neurons by hand based on the heuristics of how strong these connections "should" be. For example the connection between the text reading and speech modalities should be stronger than the connection between the colour processing and speech modalities based both on the heuristic that reading text out loud is more practised and on the fact that it takes twice as long to name a colour than to read a word out loud (Cattell, 1886). Assuming the process of learning is Hebbian, however, other connections get strengthened as a byproduct and not only those that are practised, see discussion in the section below on theoretical analysis.

The other approach is to train the CTRNN directly with the above described Hebbian learning algorithm. In this case one still needs to use the heuristics of how much training each coupling receives.

For example in the modelling of the classical Stroop task, the network was first trained to associate "colours" to "colour-words" (following the order in which human learn these things) with 300 trials each of which lasted for 100 time units and in each of which the input (I_i in Equation (1)) was equal to 1 in two neurons: one corresponding to a colour in the colour processing domain and one corresponding to the same colour in the speech production domain. In the following 1200 trials the inputs were word-word pairs (one in reading and the other in speech modality) with probability 5/6 and colour-word (again colour processing and speech production) pairs with the remaining probability 1/6. In this way the CTRNN got eventually more training associating "spoken words" with "text" than with "colours". The resulting weights are: around 3 for the connections between colour and output, around 4.8 between words and output The learning algorithm produced also non-zero weights between colours within domain and between contradictory colours which were mostly negative weights of low absolute value due to the discount factor δ in the learning rule. These were relatively randomly distributed. They had no significant effect on the performance of the network (see Results section). The weights of the connections from the task-selection neurons to the other neurons were fixed and not learned during the learning stage.

3.6.3 Theoretical Analysis

Before the results of the actual simulations are presented, I would like to present a theoretical analysis of the model which was an integral part of its design process.

MacLeod (1991) lists in an appendix "eighteen major empirical results that must be explained by any successful account of the Stroop effect". I have reproduced them in Section 3.7 and refer to them in this section to be "on the map" concerning the empirical implications of the model. I refer to them as **M1**, **M2** and so on. I have listed all of them for the sake of completeness, but some of them are not relevant for the present model. For example **M6**, **M15** and **M16** are irrelevant and cannot be accounted for: neither the spacial location of stimuli nor the age or gender of the subject are taken into account by the model. This is of course not the only source of empirical back-up, because the paper is from 1990.

One of general types of predictions that is possible to make with this model is the following: if one knows all connection strengths between different modalities, then the amount of facilitation and interference is predicted in tasks involving these modalities. Conversely, if sufficient data on these facilitations and interferences is available, the model makes a prediction about individual connections. In this way the model provides many ways to test its validity and is vulnerable to falsification.

Consider again the simplest scenario, i.e. the classical Stroop task. The modalities involved are word reading (W), colour processing (C), speaking (S) and the task control modality (T) as in Figure 4. Suppose the connections from W to S is stronger than the connection from C to S. Suppose now that the network is confronted with the task of reading the words. Then the word-neuron of T is active and so the state of each of the neurons in W is higher and more sensitive to other input. Note that the activation of that neuron is still close to zero (activation is the state to which the activation function is applied). Suppose a neutral stimulus is presented, namely one neuron in W receives a steady input. Its activation starts increasing and thereby the corresponding neuron in S starts being more and more active. Eventually one of the coordinates of μ_i will cross the threshold and output will be produced. This will take some number of time steps, say n (for *neutral*). Suppose now that a congruent colour-word pair is presented. Then the same output neuron will receive more input, but only slightly. This is because C is not pre-activated by T unlike W is and the connection from C to S is weaker, so the facilitation effect is very weak. Similarly if an incongruent stimulus is presented, then two neurons in S receive input one from W and one from C. The one that receives input from C will have the maximal activation among the neurons that *do not* receive input from W. It is obvious from equation 5 that this slows down the formation of the response, but again, not significantly. Whereas if the task is to name a colour, then the facilitation and the interference effects are stronger, because compared to the signal from C, the signal from W is stronger than in the previous scenario. The facilitation, however, is smaller than the interference, because the activation function σ is sub-additive and receiving signals a and b produces less activation than the sum of activations that would result from a and b alone and there is even a limit of how big the difference in equation 5 can be (at most 1). There is, however, no bound on how much interference there can be because the difference can be as close to zero as possible. In this way the model accounts for *the asymmetry and lower facilitation than interference.* This accounts partially for **M5**.

Each of the modalities can interpreted differently. For example if the output is not the speech modality, but button pressing modality, then the interpretation of the output set of neurons is changed from “speech” to “buttons”. Then also the weights should be altered. How is the strength of connection to the button modality (B) determined? Suppose that in the practice stage, the subject has to train pressing buttons corresponding to visual colour stimuli. Then according to Hebbian learning principle the connections from C to B become stronger. But every time a neuron in C is activated, a neuron in W is activated as well, because there is a learned connection between C and W. Thus, during the training also a W gets coupled to B, i.e. connection between the neurons corresponding to same colours gets stronger. The degree of this side effect depends on the strength of the connection from C to W. When an incongruent stimulus is then presented and the ink colour is to be named, there are two types of interference. First comes from the fact that W directly sends signals to C and the second one from the fact that W sends signals to B because of the side-effect training described above. If the connection strength from C to W and from W to C are equally strong, then this model would not predict asymmetry between word recognition and colour recognition tasks. However, if there is asymmetry, then it predicts also an asymmetry in the connection between W and C. Thus, if the classical asymmetry is present even if the output is made via button press, then this model predicts that pointing to a colour upon reading a colour word should be slower than pointing to a colour word upon seeing a colour. Whether or not this is the case, it is natural to assume that relative strength of the connection from W to B compared to C to B is not as strong as the relative strength of W to S compare do C to S predicting less interference in the button-pressing task. This accounts for the first part of **M13** and a subsequent study by (Weekes & Zaidel, 1996) where they also found that “manual responses diluted but did not abolish the Stroop interference relative to vocal responses”.

Also instead of word reading and colour processing modalities one can substitute for instance auditory and orthographic modalities. Or one can add more modalities to the network in order to analyse what happens in situations where words are e.g. semantically irrelevant but, say orthographically or auditorially, similar to an incongruent word. For example (Besner et al., 1997) report that there is interference effect by pseudohomophones such as “bloo” but it is less than the interference by normal colour words. By adding the “sound of a word” and “orthography” as separate modalities, one can analyse the strengths of mutual connections between them as well as the interference and facilitation patterns and see if the predictions match the data. This is, however, not done in the present work and is left for the future, but already accounts for **M2** since it is stated in sufficiently vague terms.

Suppose again that the output is via button press, but as in the study of van Veen and Carter (2005) two colours are mapped to the same button.

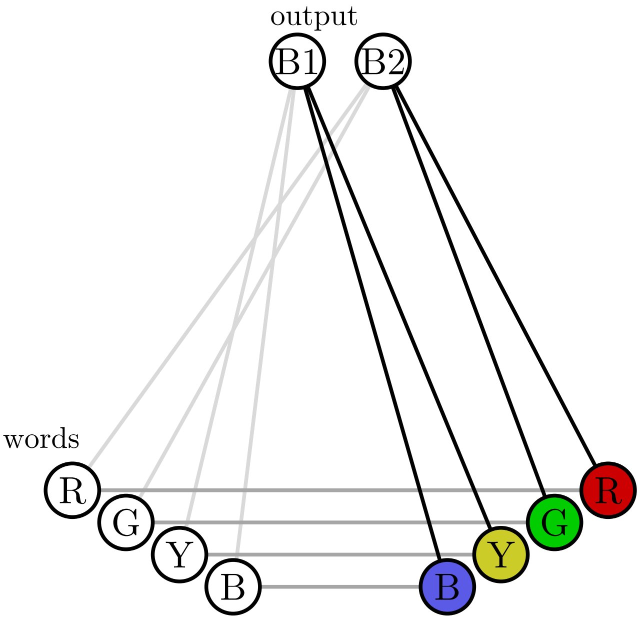


Figure 5: The architecture of the NN for the semantic incongruence versus response incongruence task.

Then the model takes the form as in Figure 5. As explained above, during the training the network will not only make a connection between colours and buttons but also between corresponding words and buttons. If the stimulus is now semantically incongruent, but response-congruent, for example the word is "blue" and the ink colour is yellow, then W will send interfering signal to C and there will be interference in the colour domain which slows down the propagation of the signal into the output modality. If the stimulus is also response incongruent, then W will send interfering signals both to C and to B thereby producing even more interference.

What about the pointing task studied by Durgin (2000)? Once a visual target has been found, pointing to it can be assumed to take a constant amount of time, so the task boils down to the visual search task: search either for the appropriate colour patch or appropriate written word. Similarly in the card-sorting paradigm (Virzi & Egeth, 1985) the time needed to execute the motor action can be taken to be independent of other factors once the appropriate box has been found through a visual search task. Now the output modality is either "visual search of written words" (VW) or "visual search of colour patches" (VC). Following the logic that "it is easier to search for the exact same thing as has just been presented", it is quite reasonable to assume that the connections from colour domain to VC and from the text-reading domain to VW are stronger than the connections from word-reading to VC or colour processing to VW. This would explain the findings of (Durgin, 2000; Virzi & Egeth, 1985) and also account for **M11**.

Above we saw how the model incorporates the clauses **M2**, **M5**, **M11** and **M13** and as also mentioned, clauses **M6**, **M15** and **M16** cannot be accounted for by the model. Let us look at the rest. Whether stimuli are presented one at a time or as a list shouldn't significantly change the performance of the model. Some information from previous trials will carry over to the following tasks, but the interference phenomenon should be robust, thus accounting for **M1**. The exact way in which the information is carried over might account for **M8**, but theoretically it does not seem plausible that **M8** would be replicated without further adjustments. In the version of the model for the classical Stroop effect there was a separate modality for task-selection whose function was to facilitate sensitivity of the task-relevant neurons. But when a neuron is actively used, it becomes on average even more sensitive, specially if the decay of the activation is not too fast (τ in Equation 1 is not too big). This would correspond to a situation where the facilitation of the task-selection modality in fact favours only those colours that are task-relevant. This would account for **M14**; the results found by Klein (1964) that the colour words denoting hues that are not in the response set interfere less than those that are in the response set. This also suggests that this model could account for **M3** and the second part of **M10** assuming that connections from concepts to colours are proportionate to the amount of "association" or "implicability"; e.g. "lemon" would be more strongly associated with yellow than "ferrari". **M4**, however, seems a little difficult to tackle. If the Hebbian learning is present not only during the training phase but also during the trials itself (it is plausible to assume that learning happens also there), then one can account for **M7**: if there are many congruent trials among the incongruent ones, then the connections from W to response get strengthened, so then the interference is also heightened. **M12** is naturally accounted for by the design of the model. Finally **M17** and the second part of **M18** can be explained in terms of the model by saying that the connections between W and any other modality are stronger in the left hemisphere than in the right hemisphere, because W is a function of the left hemisphere (see the rest of this thesis); and the connections from W to any other modality are also stronger when W stands for the dominant language rather than the non-dominant one.

3.6.4 Results

To back-up the above theoretical analysis I have conducted actual simulations with the neural networks. Below results of two such simulations are presented corresponding to two different types of Stroop task.

Standard Stroop task. As described Section 3.6.2 the network was trained with input-output pairs and was given more W-S pairs than C-S pairs (W, C and S stand for word processing, colour processing and speech modalities

	word reading	colour naming
CON	3505	3737
NEU	3946	4711
INC	4328	5444

Table 1: Reaction times of the model in the standard Stroop task presented in the raw time units corresponding to number of updates of the state of the ANN.

.

respectively). The resulting neural network was presented with congruent (CON), neutral (NEU) and incongruent (INC) trials with both tasks: word reading and colour naming. The rounded average reaction times (number of required time units for the evidence vector μ to cross the threshold, see Equation 5) are represented on Table 1.

The time units are of course arbitrary. From this we can already see the man patterns of the Stroop effect: the RT for neutral stimuli are slower for colour than for words, incongruency effect in colour naming task is bigger than facilitation and also bigger than the incongruency in the word reading task. The only feature that does not match neither the evidence nor the theoretical elaborations of the previous section is that the facilitation effect is bigger than the incongruency effect in both tasks. This problem will be addressed in the future research. But the other features do match the expectations: the neutral colour-naming is slower than neutral word reading and the interference in colour-naming task (INC-NEU=733) is greater than the interference in word-reading task (382).

Note that there was not any noise added at any stage of the model unlike in the PDP model of Cohen et al. (1990), so the model is deterministic and the reaction times will be identical every time if the weights are fixed. If I added noise, then I would have ran the experiment many times and the averages would presumably be the same as the current exact values and with enough trials the difference would have been significant; so it seems like it would have been just an unnecessary complication. In the training phase, however, the data was presented in randomised order, so the connections exhibit variation and so there is a (small) difference for example between the congruent trials red-red and yellow yellow. For this reason in Table 1. the averages are reported.

Semantic versus response interference. In the second simulation the network of Figure 5 was used. In this case the learning stage was not simulated (this is left for the future), but the weights were adjusted by hand: from colour modality (C) to button pressing (B) the weight was set to 3, from word reading (W) to C to 2 and from W to B to 1.5 following the heuristics described in previous section.

CO	SI	RI
3917	4136	4991

Table 2: Reaction times of the model in a task where two colours point to the same button allowing for congruent (CO), semantically incongruent (SI) and response incongruent (RI) trials. The interference in RI trials is larger than in SI trials.

.

As in the study of van Veen and Carter (2005), there are three possible stimuli: congruent (CO), semantically incongruent but response congruent (SI) and response incongruent (RI). The results are shown on Table 2.

3.7 Eighteen Major Empirical Results That Must be Explained by Any Successful Account of the Stroop Effect

This section is a reproduction of the appendix of MacLeod (1991).

M1. The Stroop effect is observed with lists of stimuli, with single stimuli, and with many variations on the response required. Similar data patterns are evident in numerous Stroop analogues, such as the picture-word task.

M2. Both orthographic and particularly acoustic/articulatory relations between the irrelevant word (or part of the word) and the to-be-named ink colour contribute to the interference.

M3. Compared with naming the ink colour alone, irrelevant verbal stimuli that are unrelated to the concept of colour interfere only minimally with colour naming. However, as the word's semantic association to the concept of colour increases, so does its power to interfere.

M4. A colour-unrelated word can be made to cause greater interference (or facilitation, or both) with colour naming if its meaning is activated by a related word or phrase shortly before the colour-naming trial.

M5. Congruence between the irrelevant word and the to-be-named ink colour often produces facilitation. However, this facilitation is much less than the corresponding interference in the incongruent condition, and the choice of control condition may be crucial.

M6. If the to-be-named colour and the to-be-ignored word are presented in separate spatial locations, interference will be reduced (but not eliminated) relative to the standard, integrated version of the task. Locational uncertainty makes an important contribution in non-integrated situations.

M7. The presence of congruent trials among the incongruent and control trials will tend to invoke the tactic of splitting attention over the two dimensions, thereby increasing interference on incongruent trials (Zajano & Gorman, 1986).

M8. When the irrelevant word on trial $n-1$ is the name of the target ink colour on trial n, interference with colour naming will be enhanced temporarily; when the ink colour on trial $n-1$ matches the word on trial n, there will be some facilitation of colour naming on trial n. If the word on trial $n-1$ is repeated on trial n, then the word is already suppressed and will cause less interference in naming a different ink colour on trial n.

M9. Advance cues conveying information about the upcoming Stroop trial can be used to establish processing strategies that improve performance if these cues are above the level of subjective awareness and if a very small set of cues is used consistently.

M10. When the colour (or picture) is to be named, maximal impact of a congruent or incongruent word will be observed when the two dimensions begin within 100 ms of each other. Facilitation may extend to longer SOAs than interference when the word comes first. Manipulating SOA has virtually no impact on word reading unless a very high proportion of congruent trials biases use of the colour to initiate response production.

M11. A reverse Stroop effect (i.e., interference with word reading caused by an incompatible, irrelevant ink colour) appears to be possible, but this effect is not simply a consequence of the relative speeds of processing each dimension.

M12. Degree of practice in processing each of the dimensions of a multidimensional stimulus is very influential in determining the extent of interference from one dimension on another. The greater the practice in processing a dimension, the more capable that dimension seems of influencing the processing of another dimension.

M13. Although still significant, interference (but perhaps not facilitation) is reduced when response modality is switched from oral to manual. Stimulus-response compatibility matters; if the normal processing of the irrelevant dimension leads to a response in the mode designated for the relevant dimension, interference is likely to be heightened.

M14. When the irrelevant dimension of a set of stimuli includes names that are eligible responses for the relevant dimension, more interference results than when the sets are non-overlapping. Although variations in response set size might be expected to affect interference, existing results are unclear.

M15. There are no sex differences in Stroop interference at any age.

M16. Interference begins early in the school years, rising to its highest level around Grades 2 to 3 as reading skill develops. With continued development of reading, interference declines through the adult years until approximately age 60, at which point it begins to increase again.

M17. The left hemisphere generally shows more interference than the right.

M18. Interference between the two languages of a bilingual, although not as great as that within either one of the languages, is very robust: Between language interference typically is about 75% of within-language interference. Furthermore, a dominant language has more potential for interfering than does a non-dominant one.

4 Priming Effects

One of the major themes in cognitive psychology is to understand how conscious and unconscious perceptual stimuli are processed. Which cognitive modalities and mechanisms are involved and how do they interact during word reading, colour perception, exposure to unexpected stimuli and on and on? Methodologically the question is: how to see these hidden processes? The priming paradigm provides one way to answer this question. The so-called priming effect occurs when the processing of current information is influenced by information processed shortly prior to it. In this way comparing reactions to identical stimuli but with different primes provides the scientist with a tool to see how the primes are processed. For example now you are primed to read descriptions of studies involving priming.[4]

This paradigm allows studying even unconscious processing of stimuli, because it has been found that unconscious primes influence processing of subsequent stimuli. Some of the first priming experiments were done in the 1970's. Meyer and Schvaneveldt (1975) exposed the subjects to English words ("nurse", "doctor", "bread" etc.) and non-words ("plame", "soam") and they had to decide whether the sequence of letters was actually a word. It turned out that if the word "doctor" was preceded by a semantically related word such as "nurse", then subjects were significantly faster at recognising that it was a word than if it was preceded by a semantically irrelevant word such as "bread". In this experiment subjects were fully conscious of both the prime (the first stimulus) and the target (the second stimulus). Of a particular interest to our study are the unconscious priming experiments done by Marcel (1983). One of his experiments was to show subjects a masked colour word (a noisy picture – "mask" – was presented right after the word to prevent

[4] I stole this pleasantly self-referential remark from Sternberg (2003)

awareness of the word) as a prime and a colour patch as a target. The subject had to recognise the colour of the patch. The reaction times (RT) were significantly slower for incongruent colour words than for congruent ones. This is similar to the asynchronous Stroop effect (cf. Section 8) and part of the experiment we have carried out (Section 9). This result tells that even when words are presented unconsciously, they facilitate semantic processing[5] which "spreads" to the relevant modalities, in this case colour processing. This also demonstrates that priming effect is in some sense automatic and resistant to conscious efforts. Not only are the differences in RT's so small that they can hardly be voluntarily influenced (from 50 to 200 milliseconds), but are even present upon presentation of the primes subliminally.

5 Categorical Perception: Language and Colours

The anthropologist Edward Sapir and linguist Benjamin Lee Whorf advocated the position that speakers of different languages conceptualise and perceive the world differently due to differences in grammar, usage and other linguistic differences (Koerner, 2008). This gave rise to the term "Sapir-Whorf hypothesis" which states exactly this and which is also known as the *principle of linguistic relativity*.

On the one hand this principle can be extended outside of language: people may have a variety of non-linguistic concepts in their heads which shape their understanding of the world. By restricting the attention only to immediate perception, we obtain a special case of this hypothesis. This special case avoids dealing with general, and often vague, notions such as the *world view*, *understanding* and *conceptualisation*. Is, for instance, the perception of blue colour influenced not only by the wavelength and physiology of the eye but also by the concepts possessed by the perceiver? If so, then in what way? Note the similarity in spirit to the questions posed by T. Nagel, see Section 2.1.

A lot of research on categorical perception (CP) has been done on colour categories. This isolates an easily accessible set of categories which exhibit a wide variability across languages and cultures and on which it is easy to conduct experiments and obtain statistics. The first results in this area showed that while different languages and cultures carve out different areas from the colour spectrum as "basic colour categories", all of them still agree on what are the prototypical colours of these categories (the "focal" colours) (Berlin & Kay, 1969). In this sense it seems that colour categorisation is partially cultural and partially innate. This research, however, did not investigate how exactly and why does the cultural environment and linguistic and pre-linguistic concepts influence perception. More recently experimental psychology has addressed these questions.

[5] But see Section 2.3.

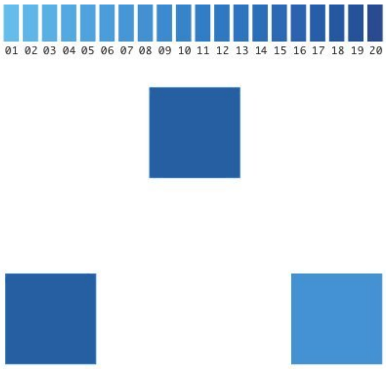


Figure 6: The participant has to decide which of the colours in the bottom row is identical to the colour at the top. Picture reproduced from (Winawer et al., 2006).

A typical experiment design is as follows. Two colours are shown to the subject one after another – call them A and B. Then colour C is shown which is equal either to A or to B. The subject has to decide as quickly as possible whether C is equal to A or to B. In a different version, the colours are shown simultaneously: A and B at the bottom of the screen on the left and right respectively and C at the top in the middle, see Figure 6. The time that is used for the decision is in this paradigm equated with the difficulty of decision. The Sapir-Whorf hypothesis predicts that if A and B belong to different linguistic categories (of a language that the subject speaks) then the decision is easier to make and when they belong to the same linguistic category, it is more difficult.

An experiment confirming this prediction is described in (Winawer et al., 2006). It is based on the observation that in Russian language there is a distinction between certain colours that does not exist in English, namely between certain shades of blue. The Russian синий /’sʲinʲɪj/ refers to a deep and dark shades of blue, while голубой /ɡəlʊ’boj/ refers light, sometimes slightly greenish shades of blue. For example the colour of the sky is often described in terms of the latter. In Figure 6 the colours on the bottom belong to different linguistic categories in Russian but not in English and it was shown by Winawer et al. (2006) that native Russian speakers were indeed faster at discriminating these colours than native English speakers. Moreover the effect could be disrupted by verbal interference which further confirms the influence of verbal categories on perception and shows that the effect is not for example due to practice in colour recognition (which would, in turn, be due to the existence of linguistic concepts).

Another way to analyse the involvement of language is to look at the lateralisation of this phenomenon, see Section 7.

The research in this area is dominated by the study of colour perception, but other category-types such as faces (Kikutani, Roberson, & Hanley, 2010) and patterns on animal fur (Goldstein & Davidoff, 2008) have been investigated as well. For a review see (Goldstone & Hendrickson, 2009).

6 Bilateral Brain

In this section I introduce, for further reference, some technical background on the anatomical and functional roles of the cerebral hemispheres.

6.1 Corpus Callosum

The corpus callosum is a bundle of myelinated axons which connects the cerebral hemispheres facilitating inter-hemispheric communication. Simple reaction time experiments where the subjects are required to press a button with a finger of the left or the right hand as soon as the light appears yield a difference in about 4 ms between contralateral and ipsilateral trials (i.e. the flash appears either on the different or the same side as the acting hand). This means that the inter-hemispheric information transfer is very rapid (Brysbaert, 1994).

In our experimental paradigm we also present stimuli in the left and the right visual field, but based on the above we cannot claim that we are testing left and right *hemispheres* per se, because the differences yielded by the priming effects are larger than 4 ms by an order of magnitude.

One might expect that due to efficiency, the stimulus is processed on the same side as to which it arrives without being sent for processing to the other side. That would be a possible interpretation of our result that verbal priming is lateralised to the left: If the stimulus is processed on the same side with verbal processing centres

In the second half of the 20:th century the research on lateralised brain function was dominated by the study of the so-called "split-brain patients", people whose corpus callosum has been surgically cut in order to treat some severe forms of epilepsy (Gazzaniga, 1970; Wolman, 2012). These people could theoretically bring insight also in the topic of this thesis, but there is only a handful of them in the world. However, in many cases as ours, if one obtains a significant lateralisation of verbal priming in healthy subjects, that is much more interesting, because one might expect that information travels back-and-forth fast enough to smooth out all these effects.

6.2 Lateralisation of Language

First evidence that language is processed in the left hemisphere came from the two patients of Broca (1861) who reported that two of his patients who had severe language production aphasias apparently related to lesions in the left

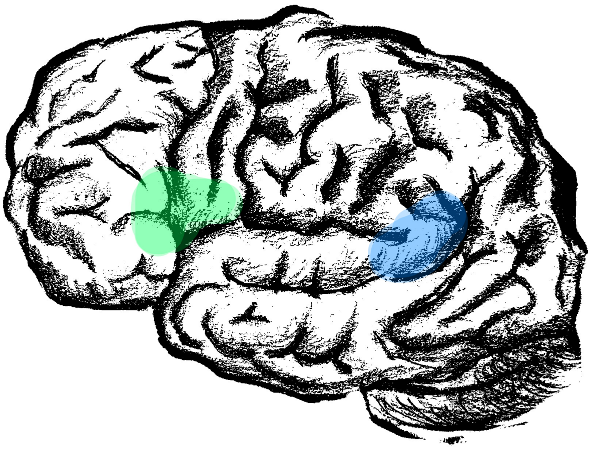


Figure 7: Areas involved in language processing in the left hemisphere. Broca's area is shown in green and Wernicke's area in blue.

posterior inferior frontal gyrus, Figure 7. This was followed by the finding by Wernicke (1874) of another area in the posterior part of the superior temporal gyrus, also in the left hemisphere, whose lesions led to semantic impairments in both language production and comprehension.

The topic of finding out the mechanisms behind language understanding and production have undergone many developments and paradigm shifts, see the introduction to (Binder et al., 1997) for a short survey. What has remained unchanged, however, is the consensus that language is processed in the left hemisphere in most people: from 96% of right-handed to 73% of left-handed (Knecht et al., 2000).

We utilise this fact in our experiment where we want to investigate the influence of language processing on priming and categorical effects. An ideal experiment could have been run on someone who has undergone a brain surgery where the hemispheres are separated through the commissurotomy of the corpus callosum (cf. Section 6.1).

6.3 Contralateral Projection

Retina at the back of the eye contains photosensitive cells which send signals along the visual pathway to the brain. Each eye outputs two streams of information, one for the left visual field (LVF) and one for the right (RVF). Because of the anatomy of the eye, the light coming from the left hits the retina on the right and vice versa, see Figure 8. The LVF is denoted by green and the

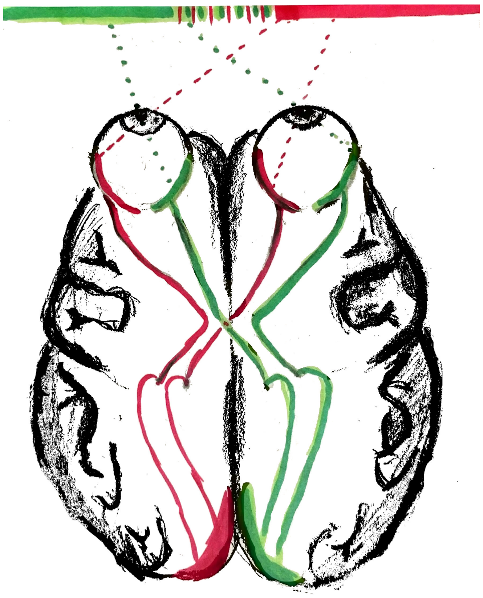


Figure 8: Contralateral projection

RVF by red and the corresponding parts in the retina are coloured with the corresponding colours. In the middle of the visual field the colours are mixed (indicated by interchange of colours), because the light coming from there hits the left retina on the left and the right retina on the right. In the figure you can see how the optic nerves from the green areas all go to the right hemisphere (RH) and all nerves from the red areas go to the left one (LH). The nasal areas (those that are closer to the nose) of the retinae send the information to the opposite hemisphere (the neural pathways cross at the optic chiasma after passing through the lateral geniculate nuclei in thalamus). The optic nerves from the temporal sides (further from the nose) do not cross hemispheres and the information is sent to the same side. The result is that all information from the green area "in the world", i.e. from the LVF is sent to the RH and the information from the RVF is sent to the LH (Purves, Augustine, & Fitzpatrick, 2004, pp. 263-267). This implies that the stimulus presented in the LVF is initially processed in the RH and vice versa.

7 Lateralisation of Categorical Perception

As discussed in the end of Section 5, the study of (Winawer et al., 2006) as well as the other research mentioned does not exclude the possibility of non-linguistic CP.

How can one distinguish between CP which is caused by linguistic concepts from that is caused by some other type of experience? Or, to put it differently, how strongly are the mechanisms responsible for CP bound to the linguistic mechanisms in the brain? Let us look at different possibilities.

FMRI. One way is to study neural basis for CP using neural imaging techniques such as functional magnetic resonance imaging (fMRI). Studies on CP using fMRI have been done on categorisation of faces (Freeman, Rule, Adams, & Ambady, 2010) and speech perception (Lee, Turkeltaub, Granger, & Raizada, 2012), but I am unaware of any studies trying to link verbal processing to CP through e.g. activations in Broca's area during categorical perception of non-verbal stimuli like colour or faces.

Infants. Another way to address this link is to study the differences between CP in pre-linguistic infants and adults. This was done in (Franklin et al., 2007) and I will get back to it below. However, there can be a significant overlap between innate and linguistic concepts and so the differences might not highlight all the valuable information.

Lateralisation of Language. Finally, one can use the fact that language is lateralised in the brain. In most people (from 96% of right-handed to 73% of left-handed) language is processed in the left hemisphere (Knecht et al., 2000).

The time frame of reaction times (RT) in the types of experiments described above is in the range of 500-1000 ms so according to (Sergent & Myers, 1985) (see Section 6.1) there is no hope in that the information wouldn't have enough time to flow across the corpus callosum. But what one may hope for is that the complex process of colour recognition and its translation to a behavioural response is happening mostly in the hemisphere to which the stimulus is presented, or at least affected by what is there in the hemisphere in question; especially under time pressure.

Due to the contralateral projection of the visual pathways, the right visual field (RVF) is connected to the left hemisphere (LH) and vice versa (Purves et al., 2004, pp. 263-267), so if a stimulus is presented in the left visual field (LVF), then it is first processed in the right hemisphere (RH) and vice versa, see Section 6.3 for details.

A typical experimental setup in addressing lateralisation of categorical perception is a visual search task. The participant is staring at a fixation marker in the middle of the screen and items appear around the marker (Figure 9). The items are all the same except one (the target) and the task is to indicate its location by moving the eyes to it. This is detected by an eye-tracker is used. The hypothesis of lateralised CP states that the RT's will be faster for items belonging to a different linguistic category than for items belonging to

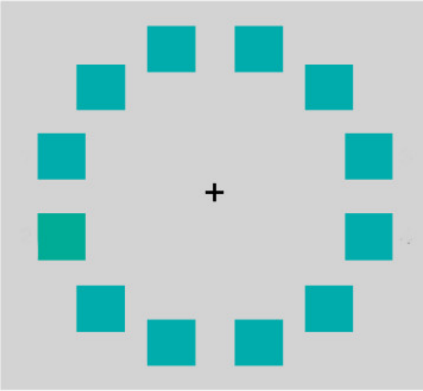


Figure 9: The experimental setup for studying the lateralisation of CP. Picture reproduced from (Zhou et al., 2010)

the same linguistic category, but only when the target is in the RVF. The first published research in this paradigm was by Gilbert, Regier, Kay, and Ivry (2006) and this study confirms the hypothesis:

> Reaction times to targets in the right visual field were faster when the target and distractor colours had different names; in contrast, reaction times to targets in the left visual field were not affected by the names of the target and distractor colours. (Gilbert et al., 2006)

Gilbert et al. (2006) hypothesised that this effect is related to the fact that language is processed in the left hemisphere. In this case, it would also confirm the presence of verbal categorical effect thereby illuminating the question stated in the beginning of this section. Note, however, that the experimental setup does not include anything verbal. The outcome of the experiment is also consistent for example with the "opposite" hypothesis that the left hemisphere is generally predisposed for categorisation and this could be the reason for language to reside there, see below for more discussion on this.

Then Franklin et al. (2007) conducted a follow-up study where it was shown that CP is lateralised to the LH in adults and to the RH in infants confirming that that the CP is indeed related to language. In a later study it was even shown that "the switch [of the lateralisation] occurs when the colour words that distinguish the relevant category boundary are learned" (Franklin et al., 2008). To test whether this effect is indeed due to learned category boundaries as opposed to innate, Zhou et al. (2010) conducted an experiment with newly learned category boundaries. The subjects who learned these new colour terms to distinguish between colours previously belonging to the same colour category showed a lateralisation effect while the control group didn't. Similar research has been done on orientation instead of colour, the categories being “vertical” and “oblique”. The advantage is that the distance between stimuli is more objective (degree of “skeweness”) as opposed to colour. Controversially they found an opposite effect than Franklin et al. (2007): the category effect for adults was stronger in the RH while for 5 months old children it was lateralised to the LH. Further Witzel and Gegenfurtner (2011) ran ten different versions of the two original experiments with overall 230 participants but didn’t find any lateralised effect. These, at least seemingly contradictory findings, raise obvious questions for further research. What is the bit of understanding that is missing from the theoretical picture?

Another question was raised by Holmes and Wolff (2012): Even if CP is lateralised to the left, it might still have nothing to do with language. They argued that the CP effect can be due to pre-verbal categorisation instead of verbal. Perhaps the LH is simply more specialised in any kind of categorisation than the RH. This in turn can be the primary reason for language processing to be lateralised to the left. They ran two experiments which confirm this hypothesis: they found a lateralised CP effect for both verbal and non-verbal newly acquired categories.

Finally I would like to refer to Al-Rasheed, Franklin, Drivonikou, and Davies (2014) where the authors show that lateralised CP does not depend on the habitual reading direction by comparing speakers of English and Arabic. Apart from the results, the paper provides a good overview of the research on (lateralised) categorical perception to date.

8 Lateralised Priming and Stroop Effects

As explained in Section 4 the priming effect paradigm can be used to infer subtle mechanisms behind perceptual processing. Can this be used to answer the questions posed in the end of Section 7? What about the Stroop effect? Can we use the knowledge about these phenomena to approach questions in categorical perception like whether the perception of colours is indeed linguistically mediated? The Stroop interference seems to suggest that it is, but as we saw in Section 3, the Stroop effect might be due to automatic evocation of the linguistic modality which wouldn’t necessarily be involved if the distractor word wasn’t present. Can we design a Stroop-type experiment where instead of the linguistic dimension we would have another colour dimension? One could present two colours at the same time next to each other and ask the participant to ignore one of them. Or, one could present them in succession and ask the participant to ignore the first one. In this case it would be a priming experiment. Since there is a lot of background knowledge on priming effects in general, we consider this is a better choice. Drawing from experience with priming (cf. Section 4) one expects that the incongruent trials will elicit longer reaction times than the congruent ones. But is this effect lateralised? If colour categorisation is indeed lateralised to the left as suggested by the experiments described in Section 7, we would expect a lateralisation also of this priming effect: a bigger interference in the left hemisphere. If the categorisation is verbally mediated and if the lateralisation is due to a connection to language and not, say, to the fact that the left hemisphere is simply better at categorising things no matter their connection to language, then we would expect a lack of lateralisation when the colours belong to the same linguistic category.

Before starting with this, however, it is advisable to see whether the standard Stroop effect is lateralised, for if it is not, then the above experiment would lose some of its theoretical power. If it is, then the next step would be to see if an asynchronous Stroop effect is lateralised: presenting a colour word first and then a colour patch. According to the hypothesis of linguistic influence on perception one would expect lateralisation in all of these cases. Fortunately some such experiments have already been conducted, although sometimes for different reasons rather than investigating theoretical underpinnings of categorical perception. Weekes and Zaidel (1996) conducted Stroop experiments with a range of procedural variations including asynchronicity and lateral presentation of the stimuli in particular to see if there are differences between the performance of different genders. In one of the experiments they would present a colour word for 150 ms and then a colour patch for 100 ms in one of the four locations on the screen: 1.5 or 4.5 degrees left or right from the fixation cross. Notice that even though the authors call it "asynchronous Stroop", this is essentially a priming study. If the prime was presented on the right side (corresponding to the left hemisphere, see Section 6.3), then the interference effect was significantly stronger than when it was presented on the left. This happened independently on the target position. We have replicated this result with a slightly different setup, see Section 9. In a review article Brown, Gore, and Pearson (1998) come to the conclusion that despite some controversy in the literature, then lateralisation of the Stroop task seems to be robust. They leave open, however, whether it is merely due to the fact that in most of the studies the responses were vocal verbalisations which are operated by the left hemisphere. The lateralisation, however, persists also with key pressing responses as demonstrated by the above discussed study of Weekes and Zaidel (1996) which is not mentioned by Brown et al. (1998) even though the former is published earlier than the latter.

9 Experiment

9.1 Design

There were 28 participants – eleven males and seventeen females between 19 and 30 years old. Each trial consisted of showing a prime for 24 ms, a pause for 120 ms and the target colour for 30 ms either in the RVF or in the LVF, see Figure 10. The participants were instructed to stare into the middle of the screen where a cross was displayed (Figure 11a). Both prime and target always appeared on the same side of the cross (either left or right), both in the same position, the offset being 6 degrees. The distance of the eyes to the screen was 55 cm. The prime was either a colour word (Figure 11b) or a coloured hash which occupies a similar space as the written word of the verbal prime (Figure 11c). The target was a colour patch of size 1.5×4 cm^2 (Figure 11d). Each participant had 30 trials from each of the possible combinations of the following binary features:

- Either left or right eye was patched,
- Prime and target appeared in the right or left visual field,
- Verbal prime or non-verbal colour prime,
- Congruent or incongruent prime-target combination.

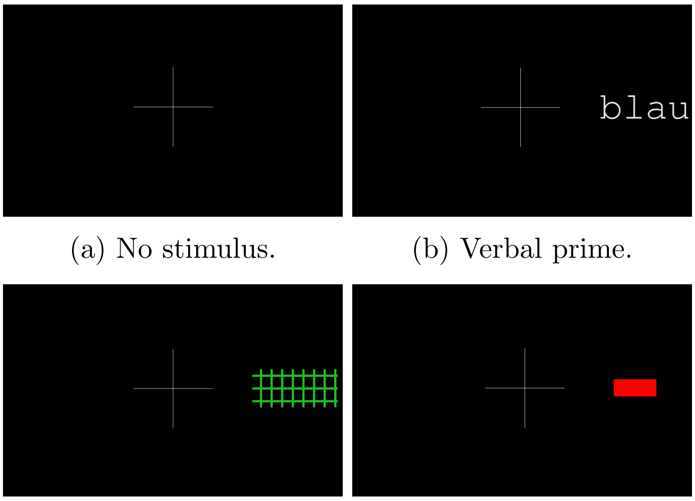
In total $2 \times 2 \times 2 \times 2 = 16$ different combinations and so each participant had $30 \cdot 16 = 480$ trials. Let us review those a little. The eyes were patched so that we could investigate the influence of the hemiretina on the reaction times motivated by findings indicating that there is an effect of this variable on priming (Huber-Huber, Grubert, Ansorge, & Eimer, 2015; Ansorge, 2003).

The experiment was run in four blocks:

1. First eye patched; colour primes,
2. First eye patched; verbal primes,
3. Second eye patched; colour primes,
4. Second eye patched; verbal primes.

24 ms	120 ms	30 ms
prime	pause	target
▦		🟨
grün		🟩
▦		🟦
gelb		🟥

Figure 10: The order and timing of the stimuli appearing on the screen.



(a) No stimulus.

(b) Verbal prime.

(c) Colour prime.

(d) Target.

Figure 11: The interface of the experiment: stimuli in the RVF.

Sometimes the first eye was the left eye and the second was the right one and sometimes vice versa. Also sometimes the dominant eye was the first eye and sometimes the non-dominant one. The first 12 participants had the blocks in the order 1, 2, 3 and 4 and the last 12 had them in the order 1, 3, 2, and 4. Within the blocks half of the trials were congruent and half of the congruent ones were on the left side as well as half of the incongruent ones were on the left side the rest being on the right. Thus 25% were congruent and on the left, 25% congruent and on the right, 25% incongruent and on the left and 25% incongruent and on the right. The order of the trials within each block was randomised.

Note that this means that the distributions of primes and targets was not independent, for if they were, we would have 25% of congruent trials and 75% of incongruent ones. In our case, the conditional probability of the target being, say, blue given blue prime was 50%. One might argue that this influences facilitation, but since priming effects are mainly unconscious and automatic (see Section 4), this effect is unlikely to be strong in comparison to the priming effect itself. This is also confirmed by the fact that verbal primes elicit much less interference or facilitation than do colour primes, see the Results section below.

The subjects were instructed to report which of the four colours (red, yellow, green and blue) the target was. They were explicitly instructed to ignore the prime and give the answers as quickly as possible. The colour was indicated by pressing one of the four buttons on a standard QWERTY keyboard: “C”, “V”, “N” and “M” standing for red, yellow, green and blue in this order. Left middle finger corresponded to red at “C”, left index finger to yellow on

"V", right index finger to green on "N" and right middle finger to blue on "M"; same for all participants except one who could use only one finger on the right hand (due to a handicap), so she used three fingers of her left hand for "C", "V" and "N" and right hand for "M". Note that if there is any delay due to contralateral stimulus, it is statistically smoothed out, because colours of stimuli and the side of their occurrence were independently randomised. Each participant had a training session lasting around five minutes prior to the experiment to learn the correspondence between colours and fingers (for the last 12 participants it was fixed to 40 trials). If the reaction time (RT) was too slow ($>$ 2000 ms), or too fast ($<$ 100 ms), then the experiment was paused until the participant wanted to continue. In case of wrong responses a sign "wrong" appeared on the screen and the experiment continued without pause. The code as well as the raw data are available from me upon request, see page 7.

9.2 Results

The data from 28 participants was first filtered such that only trials with correct answers were included and then all entries with reaction time greater than two standard deviations from the mean were ignored, the standard deviations and means being calculated task and subject specifically. In the original data there was 13373 trials out of which 12475 were correct and after the cutoff there were 11852 left, i.e. 5% was cut.

A $2 \times 2 \times 2 \times 2$ repeated measures ANOVA was conducted on this data where the factors were hemiretina (nasal; temporal), hemisphere (left; right), prime type (colour; word), and congruency (congruent; incongruent). We use the codes R for hemiretina, H for hemisphere, P for prime type and C for congruency. The main effects of both P and C were very significant ($[F(1,27) \approx 33.24; p < 10^{-5}]$ and $[F(1,27) \approx 147.01; p < 10^{11}]$ respectively) which says that, as predicted, there was an congruency effect and a difference verbal and colour priming. The two-way interactions $H \times C$ and $P \times C$ were significant ($[F(1,27) \approx 5.96; p < 0.03]$ and $[F(1,27) \approx 62.50; p < 10^{-7}]$) which indicates that the congruency effect was different depending on both the hemisphere and the prime type. The three-way interaction $R \times H \times C$ was also significant $[F(1,27) \approx 5.58; p < 0.03]$. Additionally there was a tendency for the three way interaction $R \times P \times C$ with $[F(1,27) \approx 3.44; p \approx 0.07]$ indicating the influence of the hemiretina factor on the congruency effect. The rest were insignificant ($F(1,27) < 3$ and $0.14 < p < 0.91$).

Thus, in the overall data, there is a strong dependency of the congruency effect on both the hemisphere and the prime type. We are interested whether the effect is stronger in the left hemisphere and whether this is the case for both prime types. Note that it is possible to have a significant effect in the omnibus ANOVA for the interaction $H \times C$ even if the hemisphere dependency is only in half of the data, say, word primes. Additionally the three way interaction $R \times H \times C$ suggests that there is either non-trivial role of the choice of the eye or the choice of the hemiretina.

Hence, the data was split into two parts according to the prime type and the two data sets were submitted into $2 \times 2 \times 2$ repeated measures ANOVA with factors R, H and C. For colour primes we get main significant effect of C, $[F(1,27) \approx 155.04; p < 10^{-11}]$, and a significant interaction $H \times C$, $[F(1,27) \approx 5.57; p < 0.03]$) suggesting that the effect of incongruent primes is different for different hemispheres, i.e. visual fields. No other factors or interactions were significant ($F < 3$ and $0.12 < p < 0.60$). Indeed the t-tests reveal that in left hemisphere (the right visual field) the average means for congruent and incongruent reaction times are 615 ms and 732 ms the difference being 117 ms ($p < 10^{-15}$), while in right hemisphere (left visual field) the mean reaction times are 620 ms and 710 ms the difference being 90 ms ($p < 10^{-15}$).

For word primes the same analysis yielded again a significant main effect of C, $[F \approx 23.79; p < 10^{-4}]$ and a significant three-way interaction $R \times H \times C$ with $[F \approx 4.29; p < 0.05]$. This suggests a possible difference in the priming effect across the individual hemiretinae (and not just the hemispheres) or that the hemisphere effect is prominent only in one eye. Other factors and interactions were insignificant ($F < 2$ and $0.17 < p < 0.78$). In particular the interaction $H \times C$ was not significant contrary to our expectation that the verbal priming effect would be lateralised. However, the significance of the interaction $R \times H \times C$ might reveal something even more intricate. Let us see.

If the word prime data is restricted to the right eye, then there is a slight tendency for lateralisation: the interaction $H \times C$ has $p \approx 0.12$ while restricted to the left eye there is nothing of the sort: $p > 0.98$. This suggests that the nasal retina of the right eye is the most sensitive to the primes. On the other hand restricting the word prime data to the left hemisphere, the interaction $R \times C$ is significant $[F(1,27) \approx 7; p < 0.02]$ while restricted to the right hemisphere it is not significant at all $[F \approx 0.12; p > 0.7]$. A summary of all the t-tests is given on Table 3. The above interactions and the table together suggest that the strongest priming effect is obtained in the nasal retina of the right eye, i.e. the combination of right eye with the left hemisphere, or the nasal hemiretina with the left hemisphere.

Recall that our experimental setup controlled for the dominant eye and the order of the eye-patching during the experiment both parameters being essentially randomised, so they should not have any influence on this.

Thus, we could not replicate the result of Weekes and Zaidel (1996) that there would be a significant hemisphere effect, but we see that there is a significant hemiretina effect which is reflected in the averages for hemispheres: as one can see from Table 3 (next page) the congruency effect is larger and more significant in the left hemisphere than in the right hemisphere; even though we couldn't demonstrate that this difference is significant, it is nonetheless

	Nasal retina	Temporal retina	Both retinae
Left hemisphere	46 ms, $p < 10^{-4}$	21 ms, $p < 0.03$	33 ms, $p < 10^{-5}$
Right hemisphere	21 ms, $p < 0.03$	29 ms, $p < 0.005$	25 ms, $p \approx 0.001$
Both hemispheres	34 ms, $p < 10^{-5}$	25 ms, $p < 0.001$	29 ms, $p < 10^{-8}$

Table 3: Verbal priming effects in all the four hemiretinae, and averages over hemispheres (hemiretina fixed and over hemiretinae (hemispheres fixed). In each case the priming effect is significant ($p < 0.05$). The effect is on average stronger in the nasal retinae on the one hand and in the left hemisphere on the other.

consistent with the findings that the left hemisphere is more sensitive to verbal primes and the Stroop effect in general (Belanger & Cimino, 2002; Weekes & Zaidel, 1996).

On the other hand, we also see that the strong priming effect in the nasal retina of the right eye, also apparently contributes to the fact that the priming effect is stronger in the nasal hemiretinae on average; this, in turn, is consistent with the findings that there are both attentional and prime sensitivity advantages in the nasal hemiretinae as opposed to the temporal ones (Huber-Huber et al., 2015; Ansorge, 2003).

9.3 Discussion

9.3.1 Pilot Experiment

Before conducting the experiment described above, a similar experiment was conducted which was originally supposed to be the actual experiment. There were several difficulties, however, that this experiment helped to highlight. The new experiment was designed to overcome these difficulties better.

In the pilot experiment the order of lateralisation (on which side the stimulus was presented) was not randomised and stimuli on the left were presented in blocks as well as stimuli on the right. Not only could the participants predict where the stimulus will be, but despite the instructions not to do so they also involuntary shifted their gaze towards the stimulus. In the final experiment these were randomised. Also in the pilot experiment there was less control on the distance between the head and the monitor while in the actual experiment it was fixed to 55 cm using a head rest.

In the pilot experiment I also did not try to make the colour primes resemble the word-primes (so that they consist of lines and occupy approximately the same area) and participants complained in particular that since the colours flashed very fast, if e.g. yellow was preceded by blue, it was more difficult to distinguish from green. This problem was probably not completely overcome by the new design, but should be have been better and participants didn't really complain about that.

9.3.2 Ruling Out Saccades

A saccade is a quick, simultaneous movement of both eyes between two phases of fixation (Cassin & Rubin, 2001). In the setup of our experiment it is important to take into account the possibility of saccades. If a stimulus appears in the RVF and a saccade occurs towards the stimulus so that eyes fixate at it, then it is no longer in the RVF and the information is not going exclusively to the left hemisphere. It was shown in (Saslow, 1967) that when the fixation mark is not turned off before the target onset, the saccade latency is well above 200 ms (if the mark is not turned off for 400 ms after the target onset, then it is around 250 ms). Recall that in our experiment the whole duration of a trial was 174 ms and the stimulus onset asynchrony (the time between the onsets of prime and target) was 144 ms. Brown et al. (1998) write that "Because eye movements during the Stroop display may alter the hemisphere to which visual information is projected, the display duration must be less than about 150 ms." They do not justify this number in any other way, however.

What cannot be ruled out is that if two consecutive trials occur on the same side, then a saccade caused by the first of them would cause the eyes to move towards the second trial. But since the side on which the stimulus was presented was randomised, the effect of this must be statistically smoothed out.

9.3.3 Further Research

We believe that by collecting more data, we will obtain a significant lateralisation effect of verbal priming which would probably be mostly due to it being the strongest in the contralateral visual pathway originating in the right eye's nasal hemiretina. An additional reason to believe this is that if the data is analysed as a big collection of trials (forgetting about the within subject and within task collapse to the means) then this effect is statistically significant.

Further, our hypothesis is that the lateralisation effect is due to verbal interference even in non-verbal priming. To test this, we are planning to conduct experiments with colours belonging to the same verbal category, such as different shades of red; and experiments where the verbal modality is somehow occupied: for example the subjects would be required perform verbal memory tasks simultaneously to the colour recognition task.

10 Conclusion

In this thesis three different research ideas have been presented as starting points for future developments. The first is framework theory, Section 2.3. It is a philosophical approach to a foundation for cognitive semantics and meaning. It draws essentially on multimodality, and multisensory integration. Meaning of information in one modality is evaluated based on the other modalities. The second is the Hebbian ANN model for the Stroop effect, Section 3.6.1. The Stroop effect is a multimodal phenomenon (if text reading is considered as a "modality" as it is in framework theory) and is often considered to be a way of studying semantic processing. The proposed model is in its spirit based on ideas that originate in framework theory. Finally, the third contribution is the attempt to clarify some theoretical questions in categorical perception with a Stroop-like experimental setup.

References

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (p. 339-351). Cambridge, Massachusetts: MIT Press.

Al-Rasheed, A., Franklin, A., Drivonikou, G., & Davies, I. (2014). Left hemisphere lateralization of categorical color perception among Roman and Arabic script readers. *Psychology*, *5*, 255-270.

Ansorge, U. (2003). Asymmetric influences of temporally vs. nasally presented masked visual information: Evidence for collicular contributions to nonconscious priming effects. *Brain and Cognition*, *51*(3), 317–325.

Beer, R. D. (1997). The dynamics of adaptive behavior: A research program. *Robotics and Autonomous Systems*, *20*, 257-289.

Beer, R. D., & Williams, P. L. (2014). Information processing and dynamics in minimally cognitive agents. *Cognitive Science*, *39*, 1-38.

Bekolay, T. (2011). *Learning in large-scale spiking neural networks.* University of Waterloo. (Thesis)

Belanger, H. G., & Cimino, C. R. (2002). The lateralized Stroop: A meta-analysis and its implications for models of semantic processing. *Brain and Language*, *83*, 384-402.

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

Besner, D., Stolz, J. A., & Boutilier, C. (1997). The Stroop effect and the myth of automaticity. *Psychonomic bulletin & review*, *4*(2), 221–225.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, *17*(1), 353-362.

Broca, P. (1861). Remarques sur le siége de la faculté du langage articulé; suivies d'une observation d'aphemie. *Bull Soc Anat Paris*, *6*, 330-357.

Brown, T. L., Gore, C. L., & Pearson, T. (1998). Visual half-field Stroop effects with spatial separation of words and color targets. *Brain and Language*, *63*(1), 122–142.

Brysbaert, M. (1994). Behavioral estimates of interhemispheric transmission time and the signal detection method: A reappraisal. *Perception and Psychophysics*, *56*, 479-490.

Cassin, B., & Rubin, M. (2001). *Dictionary of eye terminology.* Florida: Triad Publishing Company.

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, *11*(41), 63-65.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological review*, *97*(3), 332.

Cubek, R., Ertel, W., & Palm, G. (2015). A critical review on the symbol grounding problem as an issue of autonomous agents. In *Ki 2015: Advances in artificial intelligence* (p. 256-263). Springer International Publishing.

Durgin, F. H. (2000). The reverse Stroop effect. *Psychonomic Bulletin & Review*, *7*(1), 121–125.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429-433.

Eysenck, M. W., & Keane, M. T. (2000). *Cognitive psychology: A student's handbook.* Taylor & Francis.

Finkelstein, A., Derdikman, D., Rubin, A., Foerster, J. N., Las, L., & Ulanovsky, N. (2014). Three-dimensional head-direction coding in the bat brain. *Nature*. doi: 10.1038/nature14031

Franklin, A., Drivonikou, G. V., Bevis, L., Davies, I. R. L., Kay, P., & Regier, T. (2007). Categorical perception of color is lateralized to the right hemisphere in infants, but to the left hemisphere in adults. *Proceedings of the National Academy of Sciences*, *105*(9), 3221-3225.

Franklin, A., Drivonikou, G. V., Clifford, A., Kay, P., Regier, T., & Davies, I. R. L. (2008). Lateralization of categorical perception of color changes with color term acquisition. *Proceedings of the National Academy of Sciences*, *105*(47), 18221-18225.

Freeman, J. B., Rule, N. O., Adams, R. B. J., & Ambady, N. (2010). The neural basis of categorical face perception: Graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex*, *20*, 1314-1322.

Garland, E. L. (2012). Pain processing in the human nervous system: a selective review of nociceptive and biobehavioral pathways. *Primary Care: Clinics in Office Practice*, *39*(3), 561–571.

Gazzaniga, M. S. (1970). *The bisected brain.* New York: Appleton-Century-Crofts.

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field not the left. *Proceedings of the National Academy of Sciences*, *103*, 489 - 494.

Goldstein, J., & Davidoff, J. (2008). Categorical perception of animal patterns. *British Journal of Psychology*, *99*, 229-243.

Goldstone, R. L., & Hendrickson, A. T. (2009). *Categorical perception.* Hoboken, New Jersey: John Wiley & Sons, WIREs Cognitive Science.

Greif, S., & Siemers, B. M. (2010). Innate recognition of water bodies in echolocating bats. *Nature communications*, *1*(107), 1-6.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*, 335–346. doi: 10.1016/0167-2789(90)90087-6

Heisenberg, W. (1971). *Physics and beyond: Encounters and conversations.*

London: George Allen and Unwin.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*, 538-540.

Holmes, K. J., & Wolff, P. (2012). Does categorical perception in the left hemisphere depend on language? *Journal of Experimental Psychology*, *141*(3), 439-443.

Houlsby, N. M. T., Huszár, F., Ghassemi, M. M., Orbán, G., Wolpert, D. M., & Lengyel, M. (2013). Cognitive tomography reveals complex, task-independent mental representations. *Current Biology*, *23*, 2169-2175.

Huber-Huber, C., Grubert, A., Ansorge, U., & Eimer, M. (2015). Nasotemporal ERP differences: evidence for increased inhibition of temporal distractors. *Journal of neurophysiology*, *113*(7), 2210–2219.

Kiefer, M., & Barsalou, L. W. (2013). Grounding the human conceptual system in perception, action, and internal states. In W. Prinz, M. Beisert, & A. Herwig (Eds.), *Action science: Foundations of an emerging discipline* (pp. 381–407). Cambridge, Massachusetts: MIT Press.

Kikutani, M., Roberson, D., & Hanley, J. R. (2010). Categorical perception for unfamiliar faces: Effect of covert and overt face learning. *Psychological Science*, *21*, 865-871.

Klein, G. (1964). Semantic power measured through the interference of words with color naming. *American Journal of Psychology*, *77*, 576-588.

Knecht, S., Dräger, B., Deppe, M., Bobe, L., Lohmann, H., Flöel, A., ... Henningsen, H. (2000). Handedness and hemispheric language dominance in healthy humans. *Brain*, *123*(12), 2512–2518. doi: 10.1093/brain/123.12.2512

Koerner, E. F. K. (2008). The Sapir-Whorf hypothesis: A preliminary history and a bibliographical essay. *Journal of Linguistic Anthropology*, *2*.

Kulikov, V. (2015). Framework theory: A theory of cognitive semantics. In T. R. Besold & K.-U. Kühnberger (Eds.), *Proceedings of the workshop on neural-cognitive integration (NCI@KI2015)* (p. 8-18).

Kumar, M. (2008). *Quantum: Einstein, Bohr, and the great debate about the nature of reality.* W. W. Norton & Company, Inc.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *TRENDS in Cognitive Sciences*, *10*(7), 319–326.

Lee, Y.-S., Turkeltaub, P., Granger, R., & Raizada, R. D. S. (2012). Categorical speech processing in Broca's area: An fMRI study using multivariate pattern-based analysis. *The Journal of Neuroscience*, *32*(11), 3942-3948. doi: 10.1523/JNEUROSCI.3814-11.2012

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological bulletin*, *109*(2), 163-203.

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 126.

Maniadakis, M., & Tani, J. (2009). Acquiring rules for rules: neuro-dynamical systems account for meta-cognition. *Adaptive Behavior*, *17*(1), 58–80.

Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive psychology*, *15*(2), 197–237.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

Melcher, T., & Gruber, O. (2009). Decomposing interference during Stroop performance into different conflict factors: an event-related fmri study. *Cortex*, *45*(2), 189–200.

Meyer, D. E., & Schvaneveldt, R. W. (1975). Loci of contextual effects on visual word recognition. *Attention and performance*, 98-118.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435 - 450.

Novacek, M. J. (1992). Mammalian phylogeny: shaking the tree. *Shaking the Tree: Readings from Nature in the History of Life*, *12*, 121–125.

Peterson, J., Lanier, L. H., & Walker, H. M. (1925). Comparisons of white and negro children in certain ingenuity and speed tests. *Journal of Comparative Psychology*, *5*, 271-283.

Purves, D., Augustine, G. J., & Fitzpatrick, D. (2004). *Neuroscience* (3rd ed.). MA: Palgrave Macmillan.

Saslow, M. G. (1967). Effects of components of displacement-step stimuli upon latency for saccadic eye movement. *Journal of the Optical Society of America*, *57*(8).

Sechehaye, M. (1970). *Autobiography of a schizophrenic girl: The true story of "Renee"*. New York: Grune & Stratton Inc.

Sergent, J., & Myers, J. J. (1985). Manual, blowing, and verbal simple reactions to lateralized flashes of light in commissurotomized patients. *Perception and Psychophysics*, *37*, 571-578.

Sternberg, R. J. (2003). *Cognitive psychology.* Belmont, CA: Wadsworth Thomson Learning.

Stirling, N. (1979). Stroop interference: An input and an output phenomenon. *Quarterly Journal of Experimental Psychology*, *31*, 121-132.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.

Sun, J., & Perona, P. (2008). Where is the Sun? *Nature Neuroscience*, *1*, 183-184.

Taddeo, M., & Floridi, L. (2005). Solving the symbol grounding problem: a critical review of fifteen years of research. *Journal of Experimental & Theoretical Artificial Intelligence*, *17*(4), 419–445.

van Veen, V., & Carter, C. S. (2005). Separating semantic conflict and response conflict in the Stroop task: a functional mri study. *Neuroimage*, *27*(3), 497–504.

Virzi, R. A., & Egeth, H. E. (1985). Toward a translational model of Stroop interference. *Memory & Cognition*, *13*(4), 304–319.

Weekes, N. Y., & Zaidel, E. (1996). The effects of procedural variations on lateralized Stroop effects. *Brain and cognition*, *31*(3), 308–330.

Wernicke, C. (1874). *Der aphasische symptomenkomplex.* Cohn, Weigert, Breslau.

Wheeler, M. (2013). Martin Heidegger. *The stanford encyclopedia of philosophy (Winter 2011 edn).*. Retrieved from http://plato.stanford.edu/archives/win2011/entries/heidegger/

Wiech, K., Ploner, M., & Tracey, I. (2008). Neurocognitive aspects of pain perception. *Trends in cognitive sciences*, *12*(8), 306–313.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2006). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*, 7780-7785.

Witzel, C., & Gegenfurtner, K. R. (2011). Is there a lateralized category effect for color? *Journal of Vision*, *11*(16), 1-25.

Wolman, D. (2012). The split brain: A tale of two halves. *Nature*, *483*, 260-263.

Zajano, M. J., & Gorman, A. (1986). Stroop interference as a function of percentage of congruent items. *Perceptual and Motor Skills*, *63*(3), 1087–1096.

Zhou, K., Mob, L., Kay, P., Kwok, V. P. Y., Ip, T. N. M., & Tan, L. H. (2010). Newly trained lexical categories produce lateralized categorical perception of color. *Proceedings of the National Academy of Sciences*, *107*(22), 9974-9978.