Situational Language Understanding in Texts

A PhD Thesis

By Benedict Neurohr

Supervisors: Peter Stockwell, Kathryn Conklin

University of Nottingham, School of English

Submitted 30.09.2019, re-submitted with corrections 18.02.2021

This thesis is my own work, contains no plagiarism, and was not presented elsewhere for examination.

Disclaimer:

This thesis contains short passages and a few textual examples which were originally contained in my Master's Dissertation, *A Model of Reader Response: Genre Expectations meet Situation Model*, submitted to the RWTH Aachen, Germany, in June 2015. As this dissertation directly lead to my Thesis subject, and was in many ways its spiritual predecessor, these were arguments and in particular textual examples from Goodkind (1995; 1996) and Banks (1992) which still perfectly illustrate the points I wished to make, and I have found none better. Other sections are quoted verbatim in short passages but elaborated upon greatly and set into the new far deeper context of predictive coding, which I was entirely unaware of at the time of writing the original dissertation.

Acknowledgements:

I would like to thank both of my supervisors from the bottom of my heart for their patience and willingness to put up with me trying to squeeze two PhD projects, one theoretical and one empirical, into a single project. Peter, thank you for your optimism and steady support throughout. You gave me an opportunity to come to this incredible institution and begin something which will remain with me the rest of my life. Kathy, thank you for allowing me to come and be a part of the eye-tracking lab and taking on the job of second supervisor on my whim to work with this technology and for adopting me into the psycholinguistics research group in particular. Your advice and support throughout was invaluable, and I could never have completed this work without your input and detailed feedback on my experiments and their analyses. I am indebted to all the members of the Psycholinguistics Research Group for providing a wonderful environment of discussion, and sharing research. Similarly, I am deeply grateful to all members of the Stylistics and Discourse Analysis Reading Group which I am proud to have myself led for two semesters and the fruitful discussion there. Beyond these I am grateful to my good friend Elizabeth Stewart-Shaw who took on the task of editing a collected volume of Text World Theory research with me which became an incredible achievement for us both, and of course for having been a good friend and always having an open ear for ideas and rants of all sorts. I am similarly indebted to my friends and colleagues Fabio Parente and Dominic Thompson for many much needed tea breaks and for their vast and always openly shared expertise and advice on eye-tracking and statistics, without which this thesis could not exist. Last but not least, my eternal thanks to my incredible wife Michala, for putting up with me often barely being home, and gifting me with our wonderful daughters, Claire and Eloise. I love you all.

Abstract:

This thesis explores how human beings understand the language of fictional literary texts. The first half of the thesis explores the theory behind knowledge in the human brain, and introduces the concept of Predictive Coding and situation model theory. Using this, I present and discuss the theory which has arisen from my research on these topics, predictive model theory. Predictive model theory is used in chapter two to explore some linguistic phenomena which the theory can analyse and describe with great explanatory power. In chapter three in which I outline the main reasons why the theory is both a useful innovation to the field of literary linguistics and a powerful tool for explaining how texts can be understood in the face of ill-posed problems, fictional causality, and textual underdetermination.

Following this, I introduce my own unique eye-tracking experiment, designed to look at predictive models in real fictional text, read by readers with a state of the art eye-tracker, which approximates natural reading. In the following discussion, the surprising results which show that a version of the text which is manipulated with logical inconsistencies is in fact read faster than a control. Using my predictive model theory, I discuss this further and offer suggestions of how this gives a brand-new insight into the reading process. I then introduce the concept of contextual plausibility, and how readers use specific indicators of causality and plausibility contained within texts, and integrate these into background knowledge of the world.

In the next empirical chapter, my second eye-tracking study is introduced, which looks at how genre descriptions affect reading patterns of participants. Here I show that the unique difficulties faced by readers when part of the context of a textual extract do lead to a slowing down of the reading process. The following analysis delves further into the process of predictive situation models on a genre level.

Finally, the conclusion summarises the unique findings of my theoretical considerations and the empirical data I have gathered to support them, and how this furthers the field of linguistics.

Contents

Chapter 1: The Predictive Brain

1.1 Introduction

What does it mean to understand a text? First and foremost, it means to understand language. Language itself is a means of communication with which come many assumptions and pre-existing knowledge structures an individual has about what is being communicated. Communication as a process posits the existence of a sender and a receiver. While this is generally the case in ordinary conversation, reading literature is different. The author is absent, leading the reader to process information without being able to receive direct feedback. From a cognitive perspective, the sender is always absent as they are not present within the mind or brain of the receiver. Input from the outside world is also indirect, needing to first be filtered by our physical senses and limited by the amount of signals we are able to perceive at all, leading to the unfortunate matter that every brain indeed lives in a form of philosophical solipsism. The brain must construct what the world outside itself is like by processing signals sent to it through the central and peripheral nervous system. It does this by attributing causality to them and attempting to unravel the signals it receives into the causes, i.e. the assumed worldly objects that have caused the brain to receive them (Friston, 2002a, 2003, 2005, 2009). One of those presumed causes will of course be the sender, but whether we are face to face with the physical sender or removed by miles or centuries, our understanding of who and what the sender is remains subjective and internal. More importantly, our understanding of what a text means is based on a subjective and brain-internal process. This does not imply that all textual meanings are themselves subjective, and of course facts and logical arguments within a text certainly have objective meanings, but each time the text is read, a reader must for themselves decode and arrive at these meanings. Another individual may read a text to us, but they cannot understand the language for us. While there is good reason to believe that we, as human beings, engineer our world into a larger cognitive environment in which we encode information outside ourselves (Clark, 2015; Clark, 2011), ultimately this extended information must again be decoded and contextualised within the brain itself. The principle of communication is ostensibly to communicate meanings from one individual to the next. This process is fraught with difficulty and the conclusion must ultimately be that there is no such thing as a direct transfer of meaning from brain to brain. Instead,

there is a multi-stage process in which the brain constructs meanings, and then assumes that they apply to the world.

If predictive coding is correct, and I believe that it is and will introduce and explain this in this first chapter, then this means that the meanings constructed by our brains are always based on relevant real sensory experiences, and hypotheses about how our actual worlds interacts with our bodies and minds. My guiding question in literary linguistics in the face of this circumstance was the following: How can we understand fictional literary texts about worlds which do not exist and cannot be interacted with when our primary organ for understanding is based on our real world, and on forming pre-existing hypotheses based on prior knowledge? I am fascinated not only by this fact, but the fact that much secondary literature often discusses the way reading fiction shapes our understanding of the real world, and influences us on an emotional and intellectual level. I am far from the first to have considered this question, but I believe that I can add valuable contribution towards its answer, and I believe that all of this can be reconciled by the correct theory, as it is undoubtedly both true that fictional texts often describe objectively impossible, and even if possible then entirely fictional circumstances which we can still comprehend and relate to reality. It is a consequence of our desire to find and create meanings, or interpretations, and to match those with our sensory perceptions. In this thesis, I will introduce my theory of textual understanding, *predictive model theory,* which asserts the following: predictive coding and the principles which I will describe and apply to fictional texts based upon it can help us to meaningfully explain the processes behind fictional textual understanding, by being grounded in the situational and contextual process of reading in real time. I believe this to be a valuable contribution to the field of literary linguistics.

This thesis is guided by my attempt to answer these underlying research questions in order, building upon theoretical basics and resulting in a theory of reading, and the conclusions we may draw from it:

1. What knowledge of the brain can we utilise to describe and understand reading processes and what is predictive model theory?
2. How can predictive model theory help us to understand the systems underlying language?
3. What are the qualities and characteristics of texts which require an approach using predictive model theory in order for us to explain them adequately?

4. Using predictive model theory, what does it mean to understand a fictional text describing events which never happened and how does this happen in a typical reading process?

Our brain is an active creator of meaning rather than a passive receiver, and uses past information and context to construct meanings and ideas about the world, which are then matched to our perceptual inputs. This process can elegantly explain why and how we form entrenched opinions, quickly and efficiently interpret language, and how we can process reading fictional stories about events which never happened, and in reality could never happen. This is not a limitation, because all of our processing involves forming a hypothesis of events and then applying it to the world around us. Predictive model theory combines the cognitive theory of Predictive Coding with existing linguistic theories of language processing, most notably the concept of the situation model. The result is a flexible and context-sensitive theory which can show how human readers generally use and acquire specific textual knowledge from reading, and combine this with their existing world knowledge to form representations of what a text is describing, and continually update and improve their representations against the ongoing input as they read. My main goal was to create a framework which can show this as a general mechanism, but also operate to show how this would work with actual individuals, and the variances in interpretations gained from differing starting points in background knowledge and demeanour.

In the first chapter, I will review and introduce a body of existing literature around the human brain, and how knowledge is considered to work within it, with a particular view towards linguistics. I will then introduce predictive coding in detail, and how this can be applied to linguistic questions. I will then overview some common scenarios in which a predictive coding view can be combined with a situation model theory of thought, to form what I will call predictive models. Predictive models form the core of how we understand written language, and although other applications fall outside the scope of my thesis, all language. They are stable interpretations created by our brain to predict what it is that is being described the words we read, and tested against incoming input. I will introduce predictive model theory in detail, and the terminology which I will use for the remainder of the thesis. The first chapter will end by exploring the first principles of symbols and language which can be analysed and explained using the theory.

The second chapter will begin to use the theory in order to analyse language, beginning by looking at the idea of embodiment and symbol theories and the work of Barsalou (1999,

2003, 2009) in particular. Using predictive model theory I will argue for separate processing steps for recognising symbols and assigning specific semantic meanings to them. Following this I will discuss the phenomenon of pragmatic normalisation and how this can be explained using predictive models. In the final section I will discuss mismatch negativity effects and their relation to the different processing stages which we would expect to see in predictive coding processes in the brain.

The third chapter will turn to fictional texts, and will discuss three major characteristics of these texts which predictive model theory is suited for dealing with, and which come together to form a basis of how we understand texts. Firstly it will discuss the nature of texts as ill-posed problems, problems in which the causes for the information one is receiving and causal relationships between events are not known and for which previous knowledge and inference must be used. Secondly it will discuss the idea of causal chains, and how readers understand and also create causality in incoming sensory signals. Finally, it will discuss textual gaps, and the way in which readers fill these.

In the fourth chapter I will introduce my first empirical test of the theory, in which I ran an eye-tracking experiment using a fictional text to explore several dimensions of my predictive model theory, and how this caused readers to adapt their reading behaviour in the face of introduced difficulties. The results showed a fascinating effect: readers facing fictional texts in which the logical plausibility of events is disrupted do not slow down to process more thoroughly, but instead modify their reading and processing to speed up, entrenching more top-down processes and ultimately adopted the stance that the text was simply less rational. Readers presented with plausible if entirely fictional situations which could be resolved with some additional processing were more willing or able to expend the additional cognitive effort, and as a result were slower. These results are then used to fully introduce the concept of contextual plausibility, and how the causal mechanisms described by the text itself influence our real time interpretations of what to consider plausible or not within the context.

Following on from this discussion, chapter five will introduce the concept of genre plausibility, which is a schematic network of knowledge and expectations based on reading texts from the same or similar genres. Over time these expectations become honed and more entrenched for a given individual, and this can also nicely explain the somewhat different yet broadly convergent definitions of genre between individual readers. The discussion will frame my second experiment, in which I again used eye-tracking to study the

effects of genre plausibility in short textual extracts and to what a degree readers associated stereotypical genres with certain texts.

The sixth and final chapter will offer a summary and overview of the results of research which has gone into its thesis and of my arguments.

I will begin in the next section by giving a basic overview of the brain, and the research which I used as a basis for my research, and in turn for the basis of the concepts and research that I used in order to construct my own theoretical considerations.

1.2 Knowledge structures and the Brain

In this section and in the remainder of chapter 1, I shall begin to address my very first research question: what knowledge of the brain can we utilise to describe and understand reading processes and what is predictive model theory? In this section I will overview some of the basics of neuroscience and brain biology which formed important bases for my research, and the body of work regarding scripts, schema, and other mental processing from literature commonly already utilized in literary linguistics which all deal with our use of background knowledge. The following sections will then introduce predictive coding, and the principles of communication as interpreted through predictive coding, with a summary of the answers to my research question at the end of Section 1.5.

In order to explain what kind of background knowledge a reader might activate when faced with a text, we must first clarify what kind of background knowledge a reader may have access to, and how it is structured. The nature of expectations which humans bring to not only texts but every facet of daily life suggests that knowledge is highly structured and tends to be activated quickly. One of the first theories of how this might be achieved stems from Schank and Abelson, who spoke of scripts:

> A script is a structure that describes appropriate sequences of events in a particular context. A script is made up of slots and requirements about what can fill those slots. The structure is an interconnected whole, and what is in one slot affects what can be in another. Scripts handle stylized everyday situations. They are not subject to much change, nor do they provide the apparatus for handling totally novel situations. Thus, a script is a predetermined, stereotyped sequence of actions that defines a well-known situation. (Schank & Abelson, 1977, p. 41)

The essence of scripts is that they allow us to efficiently deal with scenarios and specific sequences of interactions which we encounter very often in daily life. While they lack

somewhat in flexibility they provide a great deal of cognitive efficiency. Once activated, a script enables us to quickly decide what is happening, predict certain likely outcomes and respond accordingly. One example of this is how we behave in restaurants by utilizing stored scripts for how they work and what we expect from them. Because of being able to simply activate the script, we do not need to learn again that we will be able to select food from a menu, that it is the waiter who needs to be informed of our choices and bring us the food and that unfortunately we will be asked to pay (Schank & Abelson, 1977). While the sequence of these events is rather strict, there is some flexibility in the form of what Schank and Abelson call 'slots.' Slots are indexical components of a script which perform a certain role but which may be filled by any number of different things. Some slots are more restricted than others and usually they contain some constraint as to what may fill them. The waiter slot in the restaurant is generally reserved for human beings of any kind, while for example the slot for colours of wedding dresses in western culture allows only white. Meanwhile the food slot may in theory be filled by any kind of food imaginable. The constraints on slots and on scripts in general are learned through repeated exposure. While some situations are common and experienced often, a script may also meet resistance when a commonly experienced situation differs from the usual. If this only occurs rarely the script will be unaffected; if it occurs with sufficient regularity then it will either become part of the original script or create a new script (Schank & Abelson, 1977).

The mental advantage of these scripts is to be able to quickly and efficiently respond to such stereotypical scenarios and to be able to relate scenarios and parts of certain scripts to each other automatically. This also allows us to infer causation through scripts: if we notice the final part of a familiar script, we would be likely to assume that the earlier parts of our script occurred and caused it (Schank & Abelson, 1977). To explore how these scripts can manifest in our heads, I will now introduce a basic overview of neuron cells, and how signals and information travel around the brain.

The brain is made up of vast numbers of neurons, and many different types on neuron cells. There is no space here to overview all of them in depth, so I shall focus on three types which are relevant and which can help to explain the processes that are relevant, which are sensory, motor and interneurons. Sensory neurons react to outside stimuli, sending information back into the brain and delivering direct input, while motor neurons relay information to and from muscles. Interneurons mediate between other neurons, forming more complex networks and either enhancing or blocking signal traffic between them (Spitzer, 2008). This is a necessary simplification of even these neuron types. Within

neuroscience, far more types of neurons are acknowledged and it is possible for many different kinds of neurons to play the part of an interneuron, such as some motor neurons which can send signals directly to other motor neurons and are defined as amacrine, or bi-directional (Squire, 2008).

Neurons in the peripheral nervous system of the body respond to stimuli from the world, sending signals to the central nervous system and ultimately the brain, where other specific neurons encode the information. Neurons are connected to each other via axons and dendrites, following the functional polarity principle: neurons send information to other neurons through their respective axon which ends in a synapse, while receiving information from their dendrites which connect to the synapses of other neurons. Interneurons behave uniquely in being able to not only send information in a single direction but also 'horizontally' to multiple neurons surrounding them at once (Squire, 2008). Neurons form connections to each other and begin to activate not as single neurons but complex networks.

The signals sent between neurons are the results of simple chemical reactions. A neuron has a specific electrical charge while resting. When stimulated the charge of the neuron rapidly drops, only to be immediately balanced by an exchange of ions within the cell membrane. This process is called 'firing' and results in charged ions travelling along the neuron's axon towards all other neurons whose dendrites are in contact with its synapse (Spitzer, 2008, p. 19). In daily life, it is simply impossible for only one stimulus to be activating one lone neuron somewhere in the brain. Scripts hinge upon the interactions which occur when multiple neurons are firing simultaneously. Synapses do not simply conduct the signal travelling along an axon but may amplify or inhibit it. Every synapse has a unique 'connection strength' which may be defined as the degree to which it amplifies or inhibits signals passing through it. If the connection strength is high, the signal is amplified and passed along; if it is low, the signal is blocked. Each neuron cell in turn has a threshold value, leading to a simple mechanism: if the signal passing through the synapse is stronger than the threshold, the neuron receiving the signal from the initially firing neuron also fires. If the signal is weaker than the threshold, the second neuron does not fire (Spitzer, 2008, pp. 21–23). Both the connection strength of synapses and threshold values together dictate whether one neuron firing is capable of causing other neurons to also fire.

The synaptic connection strength, which in essence leads to the formation of the structures called schemata and scripts above, is malleable and can be changed through experience.

That is, through specific patterns of neuronal firing. While the existing connection strength of a synapse may initially inhibit a signal, the synapse can be caused to increase its connection strength when both the neuron the synapse belongs to, and the neuron it is connected to, are caused to fire by simultaneous stimuli. This is called the Hebbian learning rule: “When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased”(Hebb, 1949; Spitzer, 2008). Every time the two neurons are caused to fire by simultaneous stimuli, the connection strength within the synapse connecting them increases. Eventually this has a significant effect: once the connection strength has been increased sufficiently through Hebbian learning for the signal from A to be above the activation threshold of B, then B will fire without its corresponding stimulus being present (Spitzer, 2008).

Hebbian learning forms an extremely important part of learning, enabling complex neuron networks to be formed by experience while also continually allowing them to be updated and changed if the corresponding stimuli are changed. In cases of extreme repetition of experiences, certain parts of these networks may become so ingrained that, as mentioned above, they will be activated without any actual stimuli being necessary. This is the mechanism that allows us to make assumptions utilizing a script or schema, such as when Schank and Abelson (1977) suggest we base inferences of causation on script structures. Hebbian learning has reinforced some scripts to the point where only one of the many possible stimuli corresponding to the script is necessary to activate the entire structure. Regarding slots, this learning process also explains the phenomena of “defaults” (Strasen, 2008). Defaults are certain values which we expect slots in a script to be filled by if we are presented with no conflicting information. We assume, for instance, that a waiter will be a human being unless specifically told otherwise, or that a wedding dress is white, wedding rings gold, etc. The logical explanation is that these values are so culturally engrained that they are experienced extremely often, potentially without any alternatives ever being experienced. This leads to these inputs being the defaults reinforced by Hebbian learning, whether they are present or not (Spitzer, 2008; Strasen, 2008). Certainly these are a powerful tool for drawing inferences and conserving cognitive processing power.

While Hebbian learning encourages positive association, the natural counter-balance to it is a phenomenon of synaptic strengths being continually weakened between neurons that are not commonly firing together. In local circuits and within individual areas of the brain this is a normal part of neuron function. While Hebbian learning encourages connections between

simultaneously firing neurons, the neurons in the immediate area which are not firing are inhibited. This process leads to local clusters of neurons having a 'winner' set of neurons which fire while suppressing all other neurons in the area. This process serves to draw stronger boundaries between associated and dissociated neurons within local clusters. It also serves to ensure that only the neurons which have the closest fit to the incoming stimuli are active, encouraging specific learning of patterns and efficient recognition of the same patterns if they are experienced again (Spitzer, 2008).

The above mechanism ensures more controlled learning and activation but requires one final step to be a useful heuristic for efficient scripts and schemata. While it is useful to define patterns by simultaneous activation, there are certain logical problems that arise when we positively associate mutually exclusive factors with the same outcome. These situations are relevant to many scripts and schemata, even deceptively simple ones. Red berries are edible if they are red but not small, or if they are small but not red; small and red ones are poisonous (Spitzer, 2008). We can use fork and knife at the Chinese restaurant if we wish, or chopsticks, but both are somewhat impractical. We can reach the UK by boat or by plane but obviously not both at the same time. These mutual exclusivities are handled by intermediate layers of neurons.

Intermediate layers are made up of interneurons, neurons which serve to mediate connections between other neurons. The vast majority of neurons in the brain appear to be such interneurons (Spitzer, 2008). In order to fully represent mutual exclusivity, also called the EXOR problem, we require interneurons. Consider an interneuron that has a synaptic connection to neurons A, B and C, but with a high activation threshold. Let us say that A has a signal strength of 1 towards both C and the interneuron. B also has a signal strength of 1 towards the interneuron and C. C has an inhibitive connection to C, such that when the interneuron is firing, C cannot also fire. C has an activation threshold of 1, while the interneuron has an activation threshold of 2. A signal coming from A is strong enough to activate C but not strong enough to activate the interneuron by itself. Similarly, B is strong enough to activate C by itself, but not the interneuron. If A and B are firing simultaneously, the interneuron's activation threshold of 2 is crossed and it fires, inhibiting C and preventing C from firing. This can be graphically represented as in Figure 1, with numbers next to arrows indicating connection strengths across synaptic connections:

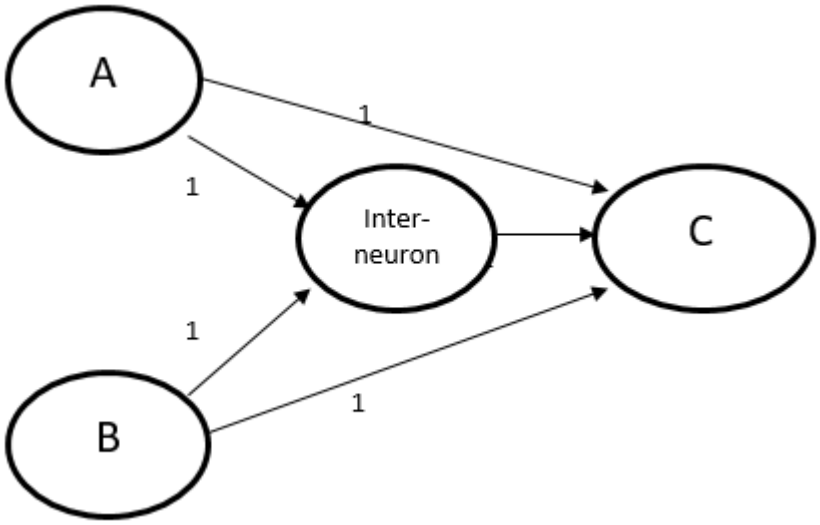

Figure 1., adapted from (Spitzer, 2008, p. 128), for Spitzer's example and a more detailed discussion, see (Spitzer, 2008, pp. 125–132).

Thus the interneuron allows for Hebbian learning to form meaningful connections, without rampantly associating any concurrent impressions with one another. Through the development of intermediary layers and Hebbian learning, scripts or schemata with highly detailed internal structures can develop through experience and habituation. Individual neurons come to form parts of many different scripts or schemata, which in turn are not physical structures in the brain but the result of synaptic wiring and connection strengths being continuously altered, forming specific activation patterns which correspond to stimuli. Intermediary layers serve to avoid inconsistencies, logical mistakes and exclusivities while also keeping the selection of factors to a minimum. In essence, this allows for the level of explanation within such networks to be minimally functional, conserving processing capacity and the need for more network complexity. As a result we tend to accept explanations at the level of the script or schema itself. This can be shown when speaking about a common winter problem of slipping on ice: Many people likely have some form of script for wearing specific footwear or walking more carefully when there is ice, for fear of slipping. Yet the situational knowledge contained in this script does not specify or contain any deeper explanation than that ice is slippery. As Sanford points out:

> Our knowledge of these particular situations means that we don't (typically) see the explanation of someone slipping on the sidewalk in winter as being due to a reduction in friction through a thin layer of water forming between a shoe or a tyre and the ice

> underneath, which leads to aquaplaning. Rather, we see the explanation as slipping on ice because it's winter. (Sanford, 2008, p. 184)

Importantly, beyond even Sanford's point here, the simple explanation suffices for someone to have a functional situational script without needing to actually know the more detailed explanation given first. In either case, given that this knowledge is learned through experience it is likely that most people would experience the event of slipping long before learning the theoretical physics underlying an explanation of friction and aquaplaning. In almost all examples of bodily experience regarding motion or the manipulation of objects, we learn the basic actions and simple causal chains long before we learn a more detailed layer of explanation. In some cases we may never learn it. I am aware that the sun will rise each morning because the earth revolves around the sun and around its own axis. I could not possibly hope to offer an explanation of the gravitational and rotational forces involved in this astounding event happening every day just as it does. Before the discovery of these forces, indeed before the discovery of the layout of the solar system, humans assumed that the sun revolved around the earth. This was not because they possessed inferior intellects but very simply because the quality of a conclusion is dependent on the quality of information available to an individual. If we do not have any further information, then an explanation such as 'it is slippery because of ice' will be sufficient in order to modify our behaviour and avoid future difficulties. Neurologically, it also allows scripts to perform their purpose efficiently by containing the necessary causal knowledge for their completion, without any excess.

Finally, while so far all mention of Hebbian learning and interneurons has been on a local scale, contained within clusters of neurons, the principles of neuronal connection can apply across the entire brain. There are long axonal connections in the nervous system and there is evidence that there can be long-range connections between different cortical areas which may also be subject to Hebbian learning mechanics (Pulvermüller, 2008). Some neurons with long axonal connections across the cortex may in turn act as interneurons for larger networks in the brain (Squire, 2008). This does not mean that script and schema structures span the entire brain as such, or that intercortex connections are as prevalent or as susceptible to Hebbian learning as the connections between clustered neurons within each cortical area. Within each sub region, neurons are ordered in tight pillars, with dense axonal and dendritic connections between the vertical layers of each column and horizontally across columns (Spitzer, 2008). Longer connections between cortical areas are necessarily composed of more isolated, long-range axons sending information between the

denser clusters of neurons. As a result, brain processes follow two inherent principles: Functional integration and functional specialization (Friston, 2002a; 2003). These principles reflect the fact that anatomically, certain regions of the brain are more closely interconnected than others, leading to a specialization for a certain task also called functional segregation; simply put, neurons with similar functional properties are grouped together (Friston, 2002a). These are also referred to as "rich-club hubs" (Park & Friston, 2013). These regions exchange signals via long-range connections leading to good reason to think that complex processes span many such hubs with detailed processing occurring within the smaller specialized hubs. Complex processing can be achieved more efficiently this way. The consequence is that it is false to view the brain as a homogenous network; processes are distributed across specialized areas of the brain (Spitzer, 2008).

Given the prevalence of interneurons and the difficulty in empirically determining how the structural connections translate into functional connections, little more can be said about the physical structure of a neuronal network and how it may represent a script or schema at this point. I do believe that each potential schema or knowledge structure is represented by a specific biological set of neurons in the brain, but as will become clear later this does not mean that there are many sets of neurons as schemata – because of the plasticity of our brains and the contextual sensitivity of our cognition, the same sets of neurons can represent many different schemata.

In this section I have argued that there is a credible foundation in neuroscience for viewing our mental processing of language in terms of functional polarity of neurons, and our prevalence to seek closed looped explanations which also aid us in understanding fictional texts. I have discussed here how information passes between neurons, that it moves in one direction between neurons but can also move tangentially from one to several other neurons and form complex networks which branch out across the brain, while still preserving the fundamental nature of information "flowing" in a particular direction - as it is never sent backwards between the same two neurons. This foundation will also help to underscore Predictive Coding. I will now present this in far more detail in the next section and present and explain all of the features of Predictive Coding and situation model theory which are necessary for my own predictive model theory which will be applied in the remainder of the thesis.

1.3 Predictive Coding and the Situation Model

In this section I will present the principles of predictive coding as found mainly in neuroscientific literature, but will also contextualise it at the end of the chapter as it is increasingly also utilised in intriguing and successful ways in linguistics and related fields as well. Following this, the section will continue by introducing the theory of the situation model, and how it works. It forms the second major part of my theoretical framework, and it will also form the bases for how predictive models are structured, and what they contain. Together, the background of this section will, together with and based on the fundamental concepts of section 1.2, form the basis for predictive model theory which I will introduce in section 1.4.

Predictive coding is based on the suggestion of Helmholtz that during perception the brain attempts to infer the causes behind an empirical signal using knowledge gained from prior perceptual experience. Helmholtz' revolutionary insight was that rather than our senses only positively adding information to our mind, existing knowledge is actually required in order to resolve what our eyes are seeing and to identify objects. As a result, we can posit that there must be an active process in the brain which at any given time applies this existing knowledge to incoming signals. (Clark, 2013; Helmholtz & Southall, 1962). With the insights modern neuroimaging has allowed, this principle has become incredibly useful for describing the way our brain approaches the world. In the study of optics and visual processing, predictive coding has been proposed as a theory capable of explaining multiple previously troublesome phenomena, such as extra-classical receptive fields (Rao & Ballard, 1999), or binocular rivalry (Hohwy, 2011; Hohwy, Roepstorff, & Friston, 2008; Jack & Hacker, 2014). I will propose that Predictive Coding, following the free-energy principle set out by Friston not only accounts for such perceptual processes but by extension also the perception and understanding of language and therefore texts. To do so, the principle will now be explained in more detail.

The brain is organized according to the principles of functional specialization and integration (Friston, 2002; 2003). As a result, certain areas of the cortex become specialized in receiving and processing specific kinds of input, such as the visual areas which are distinguishable into V1, V2 etc., the Fusiform Face Area and others. Anatomically, these areas can be called rich-clubs (Park & Friston, 2013), and they are characterized by dense arrangements of neuron cells in parallel columns with dense axonal connections between

them (Spitzer, 2008). The specialized areas are interconnected by longer axonal connections to other areas, allowing for functional integration which utilizes many smaller specialized brain areas to form and contextualize a more complex output. In Predictive Coding, this is seen as a hierarchical structure (Park & Friston, 2013; Rao & Ballard, 1999). The reason for this lies in the nature of the connections between specialized areas and within the dense rich clubs of each cortical area.

Rather than being homogenous, we can distinguish between two kinds of neuron connections running between the layers of the cortex. These are *forward connections* and *backward connections*. Forward connections typically show sparse axonal bifurcations, meaning their connections do not branch much, and run from supragranular layers to layer VI, meaning from the surface down to the deepest layer of brain tissue. Backward connections on the other hand typically show abundant axonal bifurcations, branching out heavily, and run from bilaminar or infragranular to supragranular layers, that is from the deepest most central layers of brain tissue out to the surface (Friston, 2002; 2003). The direction of the connections is significant as the functional polarity principle means the direction of an axonal connection dictates the direction of a signal sent from a given neuron (Squire, 2008). In addition, every layer of cortex has abundant lateral connections between the neurons within the layer. These make use of connections between the same levels of neurons to form representations at each level, and will also include the interneuron layers discussed in Section 1.2.

The purpose of having such separate channels for sending and receiving signals between neuron layers is a fundamental one: information efficiency and recognition. The principle of Predictive Coding states that whenever input arrives in the brain from forward connections, the purpose of backward connections is to respond based on predictions about what may have caused the input and ultimately to equalize both signals until no signals need to be sent or processed. This is a strategy which has also been discovered and used for the purposes of signal transmission and file compression in modern technology (Clark, 2013). Rather than having to process and encode the entirety of a signal, by using predictions about the nature of the final message only the difference between the signal and prediction must be transmitted. The better the prediction, the less of the signal must actually be sent and received. As we will see, this is exactly what the brain does in order to deal with the flood of signals it receives both from perception and interroception. To do so, it requires both forward connections to receive and pass forward incoming signals, and backward connections to send predictions about the nature of the signals and their causes. These

predictions are necessary not simply for the sake of efficiency, but to overcome a fundamental epistemic problem with which the brain contends.

Staying within perception, one of the principle goals of the brain is to recognize the signals it is receiving and form a representation of what is perceived. We may for now define a representation as a neuronal event that represents a cause in the sensorium, where a cause is a state of a process that generates sensory data (Friston, 2002; 2005). In this understanding, all brain processes are inherently causal. Recognition is the representation of the cause behind receiving sensory signals, i.e. recognizing that we see a couch because of the specific pattern of light exciting our optic nerves. The very serious issue the brain has in such a recognition process lies in the contextual variation of signals. Objects occlude each other, sounds become mingled and cause interference, various refractive effects in the environment can cause light to reflect and refract off surfaces unexpectedly. Language and linguistic sensory input are also highly contextual. In order to recognize the causes of these signals, the brain must attempt to undo these contextual interactions in order to arrive at a representation of the actual causes. This nonlinear unmixing of causes and context (Friston, 2002) poses a problem we may call “fundamentally ill-posed” (Friston, 2005). It is ill-posed because the brain does not know beforehand what the causes are, or what the context is, and has no access to an unaltered signal. If one object occludes another, it is impossible to receive sensory input from the occluded object.

To solve this, the brain utilizes generative or predictive models. For the sake of avoiding confusion with notions of generative grammar which are in no way related to the present discussion, the term predictive model will be used preferentially, although the neuroscience literature mostly uses the term generative model. Predictive models perform the opposite operation of recognition: they are models of causes that predict what kind of sensory signal they would lead to. These models can be learned and compared to the input in order to enable better recognition. If the brain has learned what the individual objects look like unoccluded, it can use this to model a prediction of how they would look if occluded, then compare it to the incoming sensory signal. If there is a match, the brain has found the causes it needs to represent without having had to unmix the sensory signal at all. This process is incredibly powerful, and neuronally plausible. Both modelling and representation can be equated to the nature of forward and backward connections and the hierarchical structuring of cortical areas.

The sensory input is given by forward connections which can be defined as *driving*: they terminate in the deep layers of the brain, and their postsynaptic effects are modified by the "fast" neuroreceptors AMPA and GABA which decay within 1.3-2.4 ms and 6 ms respectively. These connections quickly commit the neurons they reach to a response. Backward connections are modulatory and directly affect the synaptic effects and thus the response using the slow neuroreceptor NMDA with a decay of 50 ms (Friston, 2002; 2003; 2005). These differences in postsynaptic effects correlate to associative plasticity. Generally, when speaking of cortical connections we may distinguish between structural and synaptic plasticity. Structural plasticity refers to the way neurons are organized, dependent on physical factors of cell organization, gene expression and the development of the brain from conception. Synaptic plasticity refers to the activity-dependent formation and change of synapses (such as Hebbian learning). Associative plasticity is achieved by changing a synaptic connection to only become active if a preferred pre-synaptic effect is present at the same time as a preferred post-synaptic effect (Friston, 2003; 2005; Park & Friston, 2013). NMDA receptors used by backward connections are able to change the associative plasticity of synapses to prefer a specific response, which is then driven by a forward connection. A useful metaphor is to consider the forward connections as defining a railroad track, while backward connections control the switches and thus which direction a train (sensory signal) will take along the tracks.

The motivation of the system is to form and maintain ever better predictive and recognition models, where a recognition model is the inverse of a predictive model. Recognition models can be implemented by forward connections and can be learned via Hebbian learning. In the hierarchical architecture of the cortex, this leads forward connections to have not only a driving role, but it also constrains the signal they actually carry forward. When a signal is easily recognized by a recognition model already established, local neurons will represent the recognized causes and stop firing any further. If there is a mismatch between the signal and a recognition model, then forward connections will pass on a signal, but crucially only the mismatch will be sent on in the form of error signals (Friston, 2005; Rao & Ballard, 1999). These signals are sent by specific neurons within the network of forward connections and also have their own hierarchical ordering, mirroring that of the backward connections and the overall system. At each step of the hierarchy, error units send driving signals to higher level predicting units while receiving error signals from lower units, while predictions at each level are sent back to the error units at the same level and the level below (Friston, 2005). This is easier to represent graphically as below in Figure 2,

with units labelled EU for error unit, and PU for predictive unit. The numbers correspond to the hierarchical level of the operation. Red arrows indicate error signals being sent through forward connections to the predictive units on the same level and the level above. Black arrows indicate predictions sent back to the error units of the same level and the level below. Black arrows above error units stand for lateral connections between multiple error units, representing learned priors which help to disambiguate the signal before it is passed on (Friston, 2005).

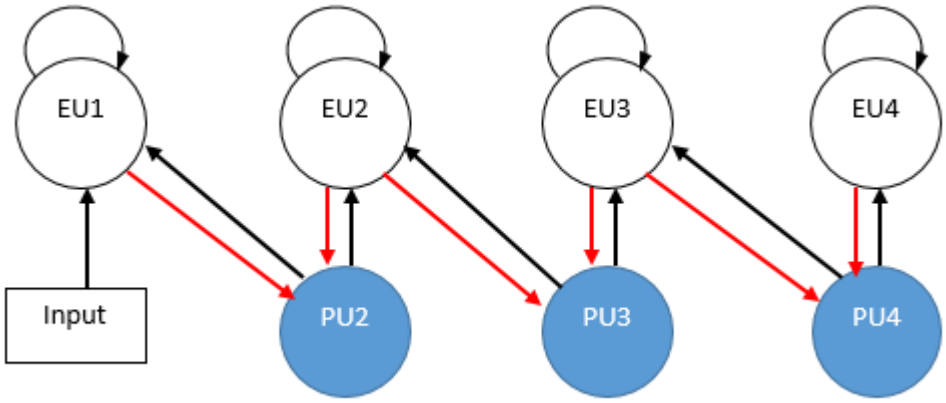

Figure 2. Error Units and Predictive Units, adapted from (Friston, 2005, p. 824).

The hierarchy of the functionally integrated cortical areas naturally comes to reflect our understanding of causes in the environment, where a mismatch of causes is explained by a new set of supraordinate causes, with each set becoming the context for the next set, and equally likely prior recognition models competing through lateral connections between error units on a given level. This naturally solves the problem of contextualizing input errors, as the hierarchy quite naturally constructs its own context out of the hierarchical organization of errors and predictions (Friston, 2005).

Ultimately, the process of recognition and understanding lies in input being weighed against the recognition models the brain has learned, sending forward errors, forming predictive models where required and sending on error signals up the hierarchy until they have been fully matched through backward connections. The principle can be defined as a form of “explaining away” what is already recognized (Clark, 2013) or in a more formal way, as surprise minimization. Mathematically, these cognitive process can be formulated along the “free-energy principle” which states that the brain as an organism attempts to minimize its entropy, by reaching a state which can represent the world with minimal neuronal

activation (Friston, 2005, 2009, 2010, 2013; Friston & Stephan, 2007; Friston, Thornton, & Clark, 2012; Karl, 2012).

So why bring this theory, which has mostly been applied to visual perception, to literary studies? First and foremost because if it is true, it tells us on a fundamental level how the human brain approaches any and all input and resulting conceptual representations, which means that it in fact applies to everything we do. Reading texts in literature, as well as hearing or telling stories, seeing films and cartoons are all perceptual processes. We are recognizing images, symbols, language. Language is in essence a perceptual system involving vision and hearing (and even touch in the case of braille for instance), evolved to represent other things, both perceptual and conceptual, but it fails to do so if it cannot be recognized perceptually first. It seems trivial, if not tautological, to say that language we do not perceptually recognize, i.e. a text we cannot read, does not perform its purpose as language to us. This does not mean that it is not language if we do not speak that language, but rather it must in some form be possible to comprehend. A script in a foreign language which I may not be able to read but which a native speaker could, is language. Random patterns on the wall which resemble nothing ever used by any group of people to communicate, is not. I believe Predictive Coding can supply a fruitful framework for analysing this perceptual process of recognising and understanding language.

While Predictive Coding has not been formally used within linguistics until now, there is a growing body of research which supports the outcomes of Predictive Coding or uses similar approaches. There are tentative links between linguistic predictions and categorisation, suggesting that our knowledge of linguistic categories and names may serve as top-down priors within Predictive Coding to shape our predictions, and what the brain will accept as resolutions to error signals (Simanova et al. 2016). Such claims are backed up by the fact that a number of studies using EEG and fMRI imaging show action centres within the brain being activated when perceiving words. Such results led to the notion of “action-perception networks”, whereby our lexical knowledge of a word is inextricably linked to knowledge of how the word should sound, how it should look when written, which hand movements are needed to write it, which mouth movement is needed to produce it, or in some cases how to perform the action the word describes semantically (Pulvermüller 2008). The action-perception network model would allow for such networks to aid in the formation of high level predictive models. Such neural networks and the principals of Predictive Coding can also be found in speech production and comprehension. Research by Cole, Jakimik and Cooper (1978, 1980) has shown that in spoken English, there are typically no boundaries

between words, as speech is produced in unbroken strings. Nevertheless, the perception that we hear pauses between individual words and segments within speech is extremely strong. Using mispronunciations in first-syllable and second- or third-syllable positions, they also showed that perception must use bottom-up and top-down processing simultaneously, as first-syllable mispronunciations are only detected after an entire word has been heard, indicating that multiple semantic meanings are active but not accepted until the perceived sounds can be assigned to one definite lexical item (Cole, Jakimik & Cooper 1978, 1980; see also Cole & Jakimik 1980). Predictive model theory will provide a framework for exactly how such a process can work and indeed agrees with their conclusions. Additionally, difficulties such as accented speech and background noise are overcome through the use of predictions and forward models (Garrod, Gambi, & Pickering 2014).

I will now introduce the theory of the situation model, and how it works. It forms the second major part of my theoretical framework, and it will also form the bases for how predictive models are structured, and what they contain. I have so far used the term "situation" several times, because it plays such an important role in my theory of predictive models. It is also widely used in the literature. Barsalou (1999, 2009) speaks of situated simulations and conceptualizations. Scripts and schemata are said to contain knowledge about experienced situations. Sanford speaks of situations and situational cognition, stating that our knowledge is organized around situations, together with the tentative definition that language descriptions are mapped to scenarios. This is assumed to be a very fast real-time process, with a successful mapping of an experience to a stored situation leading to understanding and this is also applied with minimal modification onto text understanding some years later:

> When a text is encountered, the reader is engaged in an attempt to match the text to a scenario. In the event of success, primary processing occurs, as elements of the text are mapped onto the currently active scenario. This is then used as a means of structuring the process of understanding. In the event of failure, comprehension is more difficult, and secondary processing occurs. (Sanford & Emmott, 2012, p. 24)

Situations are clearly very powerful knowledge constructs. Rather than combining the meanings of individual words to derive sentential meaning, a reader is said to retrieve a situation or many situations that constitute the basis of what is written. Consequently the skill and main objective of the writer in this view is to attempt to cause a reader to retrieve the right situations to convey this meaning (Sanford & Emmott, 2012). One result is that readers become capable of filling in informational gaps in a text by activating known situational information. Sanford and Emmott cite Schank and Abelson and their restaurant

script here (Schank & Abelson, 1970) – allowing us to conclude that in essence scenario-mapping is script theory applied to texts, including logical inferences and expected character roles (Sanford & Emmott, 2012). Scenario-mapping is also seen as the cause behind pragmatic normalization. Scenario-mapping explains this as the retrieval of a familiar scenario causing a misinterpretation by preventing a full local analysis of the sentence. It is concluded that local analysis does not precede the use of world knowledge and that readers attempt to find a relevant scenario as soon as possible during reading (Sanford & Emmott, 2012). The limits of situational knowledge are discussed, with a suggestion that: "In general, the idea is that the representations should be relevant to what people need to know in order to understand a basic situation, and no element should be more or less tightly defined than is necessary" (Sanford & Emmott, 2012, p. 34). I believe this is a very fruitful way of beginning to describe the process. A successful "mapping" is a matching of prediction to incoming signal, and the "secondary processing" in case of mismatch would be an error and prediction cycle. In the next section I shall add to this by introducing predictive model theory in order to further explore how exactly these representations might work, and be stored and retrieved as predictions while I will discuss pragmatic normalisation in more detail in chapter 2. First we must begin by exploring what situational knowledge is.

One approach attempting to constrain situational knowledge and nesting it within the context of understanding processes is that of the situation model of van Dijk and Kintsch. Van Dijk and Kintsch begin with the premises that humans construct a mental representation of something they have witnessed or are told about, and then interpret the same scenario in a certain way, such as telling a story about it (van Dijk & Kintsch, 1983). Based on these premises, they propose that the process of interpretation can be construed in the form of a real-time model of cognitive processes. The model begins with a 'textbase' which is a mental representation of the text in memory, subject to questions of coherence and which is updated in real time and under the constraints of limited memory space (van Dijk & Kintsch, 1983). The local coherence is elaborated into a connection between facts and propositions within the text, which in turn become connected into macropropositions. These can also be combined into a macrostructure, which is "the theoretical account of what we usually call the gist, the upshot, the theme, or the topic, of a text" (van Dijk & Kintsch, 1983, p. 15). The macrostructure created by these propositions is hierarchical, and based on a narrative schema and some forms of story schemata and action discourses (van Dijk & Kintsch, 1983). The macrostructure will also be updated as more information from

the textbase becomes available and is contrasted with other knowledge. The constraint of coherence together with limited memory space by necessity means that the resulting textbase is not a representation of the text as a whole, unless the text contains only a few lines. In essence, we are consciously aware of a limited portion of verbatim text which is being read at a given time, and 'relevant' fragments of what came before, combined with expectations of what is likely to follow.

The processes governing expectations are not stepwise or strictly hierarchical however:

> Again it appears that the local coherence strategies operate both bottom up and top down: Words and phrases are interpreted bottom up and fitted into the slots of strategically activated schemata—a propositional schema, frame, or script, a macroproposition, connections between propositions, expectations about probable individuals involved, and so on, all of which operate top down to provide categories or expectations about the actual information of the text. (van Dijk & Kintsch, 1983, p. 159)

In order to fit the macropropositions into such expectations, certain background knowledge is needed. The background knowledge van Dijk and Kintsch consider necessary for their model must be shaped in a flexible and responsive manner which can respond to context and to the novel situations of literature. Beyond this, while offering some discussion on Artificial Intelligence research, the question of how knowledge organization works for the purpose of the situation model is not resolved by the authors at this point (van Dijk & Kintsch, 1983). More importantly, another intermediate process is needed to govern the interaction of the textbase with this background knowledge.

This process is the actual situation model. It is a representation of the situation currently being considered by a reader, incorporating both immediate knowledge from the textbase and structured background knowledge. It forms "the cognitive representation of the events, actions, persons, and in general the situation, a text is about" (van Dijk & Kintsch, 1983, pp. 11–12). The situation model is needed as the place where the real-time processing of the interpretation of a reader is happening, and as the process which makes, amongst other things, direct reference and coreference to entities in the situation possible. When we refer to 'the man' for instance, the meaning of the noun phrase only refers to a man described in a text if we have a mental representation of him in the situation model (van Dijk & Kintsch, 1983, pp. 338–339). Once the situation model is formed it is an integrated structure of incoming information and general information from long-term memory, which must have a few key attributes: it must allow updating in real-time and it must form the basis for learning. The virtue of the situation model aside from direct reference is that it saves processing resources by only activating specific memory structures

and background knowledge relevant to the active situation model (van Dijk & Kintsch, 1983).

The description and theoretical basis of the situation model theory and its interactions with the individual text and background knowledge are in turn discussed by Strasen, who criticizes some aspects of van Dijk and Kintsch's discussion. He argues that van Dijk and Kintsch cannot offer a satisfactory explanation of how the needed background knowledge for the situation model is structured, and how only specific parts or schemata of it might be activated (cf. Strasen 2008). His other major criticism relates to the proposition that another theoretical instance, the control system, is needed as a super-ordinate mechanism to structure the situation model and to activate relevant background knowledge:

> This control system will supervise processing in short-term memory, activate and actualize needed episodic and more general semantic knowledge, provide the higher order information into which lower order information must fit, coordinate the various strategies, decide which information from short-term memory should be moved to episodic memory, activate the relevant situation models in episodic memory, guide effective search of information in long-term memory, and so on. (van Dijk & Kintsch, 1983: 12)

This leaves open how the control system should be able to perform such an impressive amount of tasks, and how the higher order information is supplied (Strasen 2008: 41). Furthermore, it seems that the control system is in danger of becoming circular: if the control system is needed to decide what information is relevant, the only place it could get such information would be from the situation model, which would in turn decide the process which selects the control system, although it already requires it to exist for there to be a situation model in the first place, and so on. For the purposes of my framework, the important tasks of this control system, chiefly providing higher order information and providing relevancy and context are fulfilled by the global model which I will explain later in this section. Despite these criticisms, the framework offered by van Dijk and Kintsch (1983), which structures interpretation as a situation modal mediating between textual input and stored background knowledge gives us a valuable theoretical tool with which to examine the structure of a possible situational representation which can reach beyond both simple logical associations and purely empirical simulations suggested by perceptual symbol theory.

In this Section I have presented predictive coding, in particular the importance of the principle of functional polarity within it. Which way information between neurons moves is important to our interpretation of empirical signals, as the signals move one way and our existing "predictions" and knowledge move in another direction, with signals effectively

nullifying each other where they meet. What crystallizes out into relevance are such situations where these information flows meet and do not match, causing errors to continue travelling along forward connections, and this is something which is critical to my own argumentation behind the processes of fictional textual. Building upon some of these studies, I will now outline predictive model theory which will be used for the remainder of the thesis.

1.4 Predictive Model Theory

In this Section, I will introduce and discuss the major concepts of predictive model theory, based on the background of literature and cognitive science introduced in sections 1.2 and 1.3. I will also begin to put it into the terms of how I intend to use the framework going forward, by splitting the situation model, or predictive model as I will refer to it, as a distinct situation within a text or conversation, from our global model, or what might be called worldview in other disciplines, which will itself be explained in more detail in this chapter. I will introduce the major concepts and definitions relevant to predictive model theory and introduce the terminology which will be used and applied for the remainder of the thesis.

Considering the sections discussed so far, it is clear that the situation model as presented by van Dijk and Kintsch (1983) offers a very useful basic framework for definition, which with minor amendments can be used within my predictive situation model theory. The textbase as set out by van Dijk and Kintsch makes sense as it stands within their theory, but it cannot be part of the situation model itself as set out in the theory presented here, as it instead corresponds perfectly to the levels of word forms and meanings. The sustained representation of the word forms taken in by the early steps of processing, together with the activation of meaning representations and the corresponding fields of knowledge in the next hierarchical step, constitute both the textbase and early macropropositions. This textbase is likely constrained just as van Dijk and Kintsch describe, with exact wording only being retained for short amounts of time, although this will vary with both experience and reader disposition; skimming a text will have drastically different textbase effects than concentrated study of a short paragraph or poem. In this sense I suggest that the textbase is analysable as separate from the situation model. It is established during the first two distinguishable levels of hierarchical processing, and it forms part of the base upon which

the higher level situation model rests. This is also fully compatible with the findings that textbase representations generally decay much faster from working memory while situation model representations stay robust for far longer (Kintsch, Welsch, Schmalhofer, & Zimny, 1990).

A useful evolution of the theory is developed under the title of the "event-indexing model," which suggests that the basic unit of the situation model is an event, described by a clause within a text. I believe the same principle to hold with verbal and visual information, although the definition of the basic unit would not be a clause per se, but an otherwise bounded section of perceived language. What this would be goes beyond the scope of this thesis, but would be an interesting question for further study. As more information is given by clauses, the event representations are added to working memory and checked for overlap with one another (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998). Overlap, or continuity, is developed across five types of event indices: temporality, spatiality, protagonist, causality and intentionality. These indices are monitored by a reader and updated when new or conflicting information about them is encountered, leading to a deactivation of a node in favour of another (Zwaan, Langston & Graesser, 1995, p. 292). With the framework of predictive hierarchical processing I am developing, it is possible to integrate this idea rather nicely. Linearity of time and space are physical constants which we learn and consequently apply to everything we perceive. Humans typically assume that space is unique, which translates into the basic impossibility for our cognition to accept that two unique objects may inhabit the same space at the same time. Causality follows the sequence of time and is a direct effect of what we look for in the world: a change in one thing is seen as the cause of another, and this also stems from the very principle of Predictive Coding, which postulates the existence of things as the very basic cause of perception. This in turn creates a chain of causality, as every change, every new perception, must be caused by something preceding it temporally, in a particular place.

Rather than talking about specific nodes, I suggest that we are talking about overlap of situation models. If a certain amount of information is already active in order to predict one aspect of a text being read, it makes sense that this information will simply stay active in order to predict more incoming perceptual information where possible. Overlap of these indices thus follows naturally from multiple situation models making use of the same underlying knowledge. The result is one of many predictive models forming a larger model

by virtue of indeed “sharing” certain activation patterns, thus naturally forming an overall activation, lending context and concurrence to one another into a single representation.

The situation model is partially constrained by the textbase, but in turn also constrains the textbase as reading progresses. Unlike the unwieldy control system, this process does not lead to a vicious cycle, but follows quite naturally from the way it is processed, via Predictive Coding. *Rather than an abstract process for everything, I would like to suggest that the situation model itself is another representation, or rather a range of representations which together make up a reader’s current model of the world being experienced – a predictive model.* The predictive model during reading is a high level representation not simply of text, but of the world as it is currently being perceived, and the world as described by the text or any alternate world described by the text. As such, it is always active in one form or another during conscious processing, sending high level predictions down the entire hierarchy. We must remember at this point that while only errors are progressively sent up the hierarchy, the final output of the predictive model representation still contains everything being received from the input – the information which did not result in errors was already active through predictions and remains active. The predictive models of all stages of the hierarchy do not cancel out the information encoded; they only ensure that this is achieved with a minimum of neural activation in the forward connections. This means all of the incoming perceptual signals must be met by predictive models, at all stages. Therefore individual word forms and meanings most certainly form part of the predictive model even at the highest level, but only once initial errors regarding their fit to the predictive model have been addressed. The predictive model also filters down, through backward connections, in turn allowing us to recognize word forms as words and as meaningful units in the first place. The predictive model can self-regulate in this way, both influencing word interpretation and being influenced by interpretation until both forward and backward connections are balanced. This means that instead of a stepwise process resulting in a “final” predictive model, the entire process as shown above is one of continuous adaptation. To be conscious means to have an active representation of the world being perceived, and this is the predictive model. The model is influenced by our perceptual data, and in the case of encountering a text, by the language in it, plus context. The conviction that there is a world, and we are interacting with it is always passively present as part of the world model that our brain itself becomes. Picking up a text and reading it leads to an adaptation of the predictive model to integrate that there is an object, that object is a text, that it contains symbols, that these symbols are

meaningful, what the meaning is, and how the meaning fits into a representation of a state of affairs described.

This does not mean that it is a representation of the entire world, which is impossible as our senses cannot perceive the world in its entirety, or even of the entire knowledge an individual might have about the world. Rather, this representation governs the world immediately being experienced, bounded by the limits of perception. This means it includes whatever an individual currently perceives within the range of their vision, hearing, reach of tactile perception etc. While it is indeed the case that an individual's knowledge about the world as a whole is used in order to form optimal predictive models in processing, this knowledge does not need to be activated as a whole, or even as a fully-fledged representation in its own right. Rather, as suggested within Predictive Coding a model of the world per se is not necessary in cognitive processing because the neural connections themselves come to stand for such a model. Our beliefs and our representations of self-agency as well as our understanding of linguistic representations require a top down model of the world against which meanings are compared. This model must be flexible and must be usable as comparison regardless of its validity (or perceived validity). Without this, imagination, daydreaming, counterfactual reasoning, learning and fiction would not be possible. This exact model is what the brain itself, as an amalgamation of all inputs, becomes as it learns from the input received from the world. Whenever a predictive model needs to be formed in order to process current experience, this basic background model embodied in actual neuronal connection strengths gives the primary top down predictive feedback of how things should be, or to be more precise of *how an individual believes things should be based on convictions and prior experience*.

I shall speak of this overall, highest level representation of the world which is passively present in our brain at all times as the "global model." This is the sum of our knowledge of reality and of our remembered conscious experience of it. By necessity, some of the basic things we know or believe about the world from our global model therefore inform our basic interpretation of words. Our representation of the situation, in its entirety, is compared to the information given by the sentence and all possible interpretations of it. The best fit is selected. This leads to the kind of error giving rise to pragmatic normalization, and other errors of processing. If the sentence is internally flawed, that is if it contains an error of syntax or semantics within itself, an error at that level of the hierarchy is elicited. If the sentence is internally acceptable, but clashes with the predictive model because it cannot be integrated into the global model, then a higher level error occurs and the

sentence meaning will not be accepted at face value. Instead, a new predictive model will be formed to attempt to deal with this error. Either the predictive model must be amended, or the sentence. One result would be conceptual learning, another to accuse a speaker of having made an error, or simply implicitly assuming this and internally correcting the perceived error. The individual representations of sentences in turn are assimilated into a larger representation, which dictates the overall plausibility and coherence of the representations, and this in itself forms a process which self-monitors through errors and predictions at the predictive model level.

The kind of predictive models evoked by texts and individual experiences at any given time are "full predictive models" or more simply, predictive models. There is one smaller unit of representation, which corresponds to a minimal amount of information needed in order to form a full representation which may be tested against a prediction, which I shall call "minimal predictive models." The relationship between them is such that minimal predictive models feed into the predictive model as individual event chains, or situations, which are resolved into a larger structure of "what is happening." This is similar to Sanford and Emmotts scenario mapping as mentioned in section 1.3, repeated below for convenience, but we can now fully contextualise the major difficulty faced by it.

> When a text is encountered, the reader is engaged in an attempt to match the text to a scenario. In the event of success, primary processing occurs, as elements of the text are mapped onto the currently active scenario. This is then used as a means of structuring the process of understanding. In the event of failure, comprehension is more difficult, and secondary processing occurs. (Sanford & Emmott, 2012, p. 24)

The problem faced by this, is that we have no original heuristic for delineating "event," "situation" or "scenario." There is much talk of individuals learning and recognizing them, but not how or why this should be a basic conceptual currency. There is also an unaddressed circularity in the fact that individuals are supposed to be able to relate new experiences to past experiences without explaining how they learned the past experiences before having an original concept of what was happening. There is neither room for learning situations the first time around, nor for learning entirely novel ones.

Predictive model theory can address this by saying that we are not comparing new input to an unspecified amount of previously stored scenarios, but rather to the already active predictive model. *There is in fact one single prior concept in the human brain of what an event, situation, scenario, happening and so forth is, which is represented by a predictive model structure. Situations are differentiated from one another not by a change in the fundamental structure of what a situation is, but only in the values of variables they*

contain. A situation is like a logical equation with set terms which are variables that can add up to an outcome of either acceptable or unacceptable. Acceptable and unacceptable may be defined in terms of Predictive Coding. *Whenever a predictive model can be integrated with both the bottom up sensory signals coming in and the top down predictions filtering down from the global model without remaining error, it is acceptable. Whenever errors remain which force the predictive model to be changed in order to integrate it without further errors, it is unacceptable*. I strongly refrain here from calling these outcomes right or wrong, or true or false. The reason for this is very simply that an acceptable outcome for the individual has nothing to do with truth or falseness, but only with the context of the overall knowledge and beliefs available in the global model, and the quality and definiteness of the available sensory input. Acceptable, and "true" in the purely subjective sense of what the individual believes to be true, means that the individual's predictions can satisfactorily match the input signal it is perceiving and no more. I will discuss problems with referring to this circumstance as "truth" in far more detail in chapter 3, section 3.2. The basic structure of the minimal situation follows from the very basic beliefs evident in all human writing and expression, and it follows what I shall call *the one-to-one principle: one agent performs one action at one point in time at one spatial location. This fundamental structure can be elaborated to contain more objects, such as a patient for an action, more actions, object attributes, or a change in attributes, and more complex situations are simply chains of this minimal situation.*

Even if not all parts of the minimal situation are explicitly mentioned in an utterance, they are implicitly assumed by the formed representation of a reader. These assumptions are only noticed if there is a conflict between them, and to give a very short example we can see these expectations at work when giving conflicting statements where the conflict is only implicit. Consider being told "David finished his homework and took a shower at 5:30 pm," versus being told "David was on a flight to Chicago and asleep at 5:30 pm." We should be very surprised to be told that in the first sentence both actions were actually true at exactly 5:30 pm, but not so for the second sentence. This is because we assumed those actions to have particular requirements of time and place, and we know from our global model that taking a shower and doing homework have mutually exclusive time and space constraints, while being passenger on a plane and sleeping do not. The natural way to integrate this into a predictive model is to represent a minimal predictive model in which David takes the shower at 5:30, and another minimal predictive model of him finishing his homework just before 5:30, even though this sequence of events is not given by the

sentence. This seems like a trivial example but we will see that even very complex situations always lead to such assumptions and can greatly confuse readers of certain texts. The same circumstance is however also responsible for coherence between representations and minimal predictive models in order to form more complex full predictive models which become integrated into the global model.

Returning to the ideas of Zwaan et al. (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998), the idea of certain core parts of event representations forming indices around which multiple representations can become organized fits neatly into the model I am proposing as well. The incoming signal of reading a text runs along forward connections and is driving during processing, meaning that the brain is forced to respond to and interpret these signals. Predictive models must account for this, by either anticipating the driving signal, or adapting to it. Following the free energy principle of Friston (Friston, 2005, 2009, 2010, 2012, 2013; Friston & Stephan, 2007; Friston et al., 2012), the preferred state for the brain is when both forward and backward connections are synched to activate the same neurons, resulting in what I define as the "acceptable" state of a predictive model. If a prediction immediately fits an incoming signal well, then this is the default state already, and all the backward connections must do is to inhibit any other associated neurons that might also fit the forward signal, but do not match predictions. Note that this does not represent an all-or-nothing representation: it is entirely possible for multiple interpretations of the forward signal to remain active, as long as the predictions also contain all multiple interpretations. The brain naturally becomes organised such that forward and backward signals overlap as much as possible, with any remainder becoming an error to be resolved by a higher level of the hierarchy. This is how it is possible to consider multiple interpretations to be true until further information can be obtained.

As a text is read, all the new information is added not simply as entirely new and unique input, but because of the full predictive model in play, a reader is aware that the text is meant to represent a continuous description and that sentences will be related to each other. Taking this piece by piece, let us say that the very first proposition of a text leads to formation of a minimal predictive model. This minimal predictive model sets up the first index values for the involved objects, actions, the time and place. As more sentences and clauses follow (or sounds, in verbal discourse), new information is added to the minimal predictive model, forming a full predictive model. All the information from the combined sentences is represented in this full predictive model. Now it follows quite naturally that if

the value for any of the indices, say the temporal value, is the same as in the previous sentence, then by virtue of the full predictive model the neuronal populations representing this value are already active. The same for indices of spatial location, causal chains, agents and goals. Unless the values of these indices is explicitly or implicitly changed, the full predictive model, mediated through the modulated plasticity of backward connections will simply keep the patterns of activation constant to minimize overall activation and free energy. Thus we naturally assume until told otherwise that a chain of events is occurring at the same time and place, with the same agents and with one action forming a clean causal chain to another. The causal chains themselves form naturally as further predictions need to be matched to new minimal predictive models. In this way, predictions begin to not only predict, but themselves become representations of the causes of the input, whether they truly are the causes or not. The hierarchical structure of predictions naturally forms a hierarchical structure of causal chains as each validated prediction becomes a basic cause for the next higher prediction.

Most importantly, as causal chains and as representational chains, multiple minimal predictive models and full predictive models can be represented at once and be associated to the global model to varying degrees. As such the actual content of a full predictive model when contrasted to the global model during the reading of a fictional text includes not just the content of the text, but something like “Being in the real world at location X reading text Y which talks about events Z.” Within the representation of this global status, multiple minimal predictive models about the actual world location, information of other things being noticed in the background, as well as background knowledge about the text in question will be present. Selective attention will modulate to which a degree each of these representations are in focus. Some may also be combined, or blended, following the work of Turner (1998) which integrates well into the current theory and can be interpreted as combinations of predictions forming hybrid representations based on multiple stored priors. Ultimately, as the highest level of the hierarchy, the global model is the basis against which all input is compared and forms the basis against which plausibility is calculated, and what is considered an error. Every representation, from the individual symbol up must be able to be integrated into it, but also has its recognition model validated by it. We recognize symbols because it is part of our global model that the world contains meaningful symbols, and in turn the global model comes to contain the interpreted meanings of the perceived symbols mediated by all intermediary steps. This will be discussed in more detail in the following section. This same circumstances can also cause conflicts when incoming input

contains information that cannot be easily integrated. *No matter how good the predictions at the word and sentence level can fit a phrase, if the content of the phrase as a minimal or full predictive model cannot be integrated into the global model, the minimal or full situation must be amended by further predictions or input until it can. Where there is an unresolvable conflict, the highest level prediction which is coherent with the global model will be preferred even if errors at lower levels, such as at the individual symbol or word meaning level, remain.*

The situational overlap discussed by Zwaan et al. (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998), is this exact principle, which naturally emerges from the way predictive models attempt to integrate new input into the minimal or full predictive model and ultimately the global model. This overlap also naturally creates internal coherence and allows complex representations of a "situation" which increase in complexity as more information is added around the same core index values of time, location and principal objects and agents involved. In this way we can see multiple hierarchical levels within predictive model theory. The way in which this hierarchy of minimal and full predictive models and the global model interact is represented in Figure 3. As new input travels up forward connections and meets previous predictions, minimal predictive models are formed, and as remaining errors from these propagate upwards full predictive models are formed which are in turn integrated to the global model, from which any remaining errors cascade down. As we will come to see throughout this thesis, not all incoming signals may ever need to go all the way through the hierarchy, and many everyday scenarios and interactions in which we are well versed will be integrated easily.

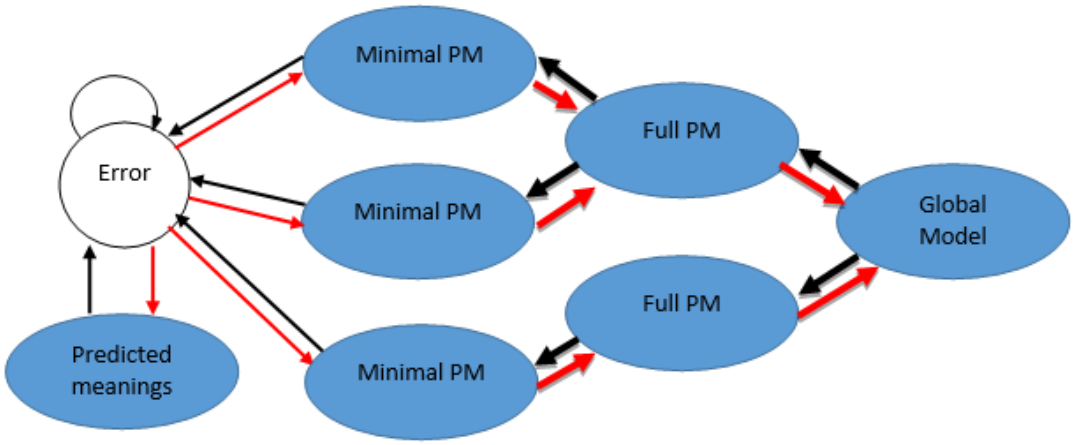

Figure 3. The hierarchy and interdependence of input, predictive models and global model

We can now see that indeed many of the definitions of “situations” or “scenarios” as described in the literature follow very different stages of complexity, sometimes describing a minimal predictive model, sometimes a full predictive model and sometimes even the global model as a whole. The situation model as discussed by van Dijk and Kintsch (1983) corresponds to an unspecified combination of the global model and a full predictive model, with the textbase forming minimal predictive models and macropropositions being full predictive models about the text. Zwaan et al. (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998) generally discuss full predictive models and the integration of new clauses.

Finally, predictive model theory encompasses two further principles when reading texts: contextual plausibility and genre plausibility. These are unique expectations and pieces of information which can be used by a reader during online reading to uniquely explain and suppress errors by utilising information given by a text itself while reading, or from past experience of reading particular types of texts and genres. These will be explained in far more detail in chapters 4 and 5 alongside the empirical experiments which led me to include these concepts in the theory.

In this chapter I have introduced predictive model theory. The predictive model is the description of our real-time representations on what is happening around us at present. When reading, the current empirical signals of the text being read form a textbase, which top down predictions attempt to match and neutralise, or explain. Minimal predictive models convey minimal event structures based on the one-to-one principle: a single agent taking a single action at a single point in time and place. Others form macropropositons and expectations of the kind of text it is and how it may unfold. I have also argued that these are compared at the highest level to the global model. The global model is our worldview – the aggregate knowledge of the world as experienced by an individual and our concept of the world against which we compare new information. As input is received, minimal predictive models are formed into full predictive models which are continuously fed further input from forward connections, and integrated to the global model through backward connections, attempting to bring both into balance. In the next section I will begin to discuss how this theory can be applied for analysing language, before moving on apply it in topics originating in pure linguistics in chapter 2, and to the study of texts in particular in chapters 3, 4 and 5.

1.5 Predictive Model Theory and language

In this Section I will begin to relate predictive model theory to what I believe are particular traits of human communication and language which pose much theoretical difficulty, and which predictive model theory can begin to analyse and explain. I will present and discuss the Chinese Room, a particular thought experiment by Searle (1980), and discuss how this can nicely show how predictive coding and predictive model theory work to make sense of symbols in our brain, and the necessary distinctions to be made between symbols as a system unto themselves and their semantic meanings as another system.

I shall now begin to discuss why it makes sense to describe language comprehension in terms of perceived input, predictive models, and meaning lying in the successful integration of bottom-up perception and top-down prediction, keeping in mind at all times that this separation is only on the lowest hierarchical levels, while at the higher levels, all of these processes become integrated into one complex process of understanding and forming language. I will make no attempt to locate any place that any word or meaning might live in the brain; partly because I do not know where they might be, but more importantly because it does not matter to the current endeavour. The beauty of the human brain and adaptive plasticity is that any area can, when needed, become specialized to deal with any form of input, then connect to any other area in order to contextualize its operations. Likely many actual neural networks are subtly different in each individual, owing to variations in the epigenetic makeup of their brain, their lifetime patterns of input and experiences and their particular strategies of predictive models. Many of these structures will be similar but unique to the individual. *What is not unique to the individual is the mechanism behind it all*. That mechanism is what I am describing in this thesis, and the empirical tests which will be discussed in chapters 4 and 5 are a beginning at discovering the effects of the mechanism in real-time reading processes. A useful first step is to discuss the nature of linguistic symbols in the brain and how they might be processed through predictive models.

The main issue that has plagued theories of symbols is the symbol grounding problem (cf. de Vega, Glenberg & Graesser, 2008), which is a response to many theories of computational linguistics. Symbols, such as words, are held to be arbitrary, abstract and amodal (not tied to any particular sensory modality), leaving open the question of how we as humans "ground" or ultimately justify connecting meanings from our consciousness to these symbols and the world. A similar problem is discussed by Searle in the "Chinese room argument" (Searle, 1980). The problem at hand is a thought experiment: Searle is sitting in

a room and is fed slips of paper filled with Chinese writing. He does not speak Chinese and cannot read them. He is then given a second type of slip with Chinese writing, but which also includes English rules for how to correlate symbols from the first batch to the second. Finally he is given a third slip, which contains more symbols and English rules for drawing out certain symbols he gains from comparing all three slips. He dutifully draws them, and returns them to whomever may be out there feeding him the slips. To the outside world, it would appear as if Searle speaks perfect Chinese, while he in fact has no idea what he has read or responded (we shall play along with Searle and not ask why they would have felt the need to include English instructions if the box inhabitant seems to understand Chinese) (Searle, 1980). The fundamental difficulty Searle is trying to showcase is the same as the symbol grounding problem: how can the human mind connect mental structures to physical structures in the world so that we do not just manipulate symbols but understand Chinese? (de Vega, Glenberg, & Graesser, 2008). The answer of de Vega, Glenberg and Graesser as well as others is to consider symbols to be inherently embodied:

> Linguistics symbols are embodied to the extent that: (a) the meaning of the symbol (the interpretant) to the agent depends on activity in systems also used for perception, action, and emotion, and (b) reasoning about meaning, including combinatorial processes of sentence understanding, requires use of those systems. (de Vega, Glenberg & Graesser., 2008, p. 4)

The natural elaboration of this definition into an entire theory of embodied symbolic representations is that of Barsalou and his perceptual symbol theory (Barsalou, 1999). Standing as a counter-point to amodal symbols, Barsalou proposes that symbols are perceptual and modal. Schema theories and many others are, as Barsalou laments, amodal. He believes that this is a weakness of the theories, as they only show arbitrary connection between symbols and the perceptual states that correspond to them. Symbols and perception are considered to be parts of different processes within the brain, underlined by separating cognition from perception (Barsalou, 1999). Instead of this, he suggests we take a more direct approach, combining the two into one. Perception can directly produce symbols within the brain which are modal and not arbitrary. They must however, overcome some principal problems which arise if it were simply a literal translation of photographic perception into cognitive “photographs.” Such a view cannot explain abstraction, or the possibility of productively taking parts of past perception and combining them into novel thought processes (Barsalou, 1999). These weaknesses do not apply to the right perceptual theory, Barsalou insists. Instead of entire perceptual states, only small subsets which represent a coherent aspect of them are preserved. These states are encoded through selective attention, prioritising these coherent aspects and storing them in memory

(Barsalou, 1999). Thus, he defines such symbols as an “associative pattern of neurons” (Barsalou, 1999, p. 584). They can be activated dynamically in the future in order to simulate past symbols, lone objects or even event chains in thought. Simulators are the main goal of learning in this view, allowing abstraction, understanding and potentially infinite imaginary variations of a theme (Barsalou, 1999). In essence, this theory seems to be based on very similar principles of learning as I have outlined in Section 1.2, without citing them in detail or describing the cognitive processes involved in such a thing. The idea of dynamic activation and simulation as an imaginative process is an intriguing one, which merits further regard.

Barsalou himself elaborates upon the idea of simulation by adding the concept of situatedness to it. During understanding, the perceptual symbols stored according to the theory are not just stored in isolation but with a context, or a background:

> Three factors dynamically determine the most accessible subset of a concept’s content on a given occasion: frequency, recency, and context. As information is processed frequently for a category, it becomes better established in memory, and more accessible across contexts. (Barsalou, 2003, p. 545)

The process of learning concepts and perceptual symbols now rests completely on the same cognitive principles outlined in the previous section, while adding a much needed component of context. These concepts he calls “situated conceptualizations” which also contain likely actions and states an agent might have in the pursuit of a goal (Barsalou, 2003). This also adds a component of prediction to these conceptualizations, as Barsalou suggests stored conceptualizations can also be utilized in order to assess the current situation and quickly fill in needed inferences. The concept is mentally simulated and used in order to decide on an optimal course of action (Barsalou, 2009).

In many ways, especially the later outline of a situated conceptualization is a more detailed and refined reformulation of Schank and Abelson (1977). Experience becomes perception, both are still stored in memory, both provide chains of actions based upon neural activation, both are activated within a specific context. The fundamental difference, the one that makes Barsalou’s account incredibly interesting, is the idea that the conceptual structure can be wilfully manipulated by the mind in order to create new conceptualizations, by utilizing the same cognitive apparatus used to take in the perceptual data in the first place. There is a very strong interpretation which follows from this claim, along with a weaker but equally productive interpretation. The strong interpretation is that Barsalou is placing the act of imagination firmly within the perceptual centres of the brain, which enact a pseudo-perception, a kind of perception without a stimulus in order for us to

creatively imagine. The other entailment of the strong interpretation is that our imagination is indeed ultimately limited by our perceptual experience: while we may productively re-combine and mix around aspects and selective components within a simulator, these components are always based upon the neurological patterns stored from an original perceptual experience. Even if they are dynamically activated, the source of the components would always be empirical. I shall instead interpret Barsalou more weakly, as saying that there is a mechanism, however it may work, which simulates a current situational state in the brain, making predictions about how it may change or freely manipulating it within an imaginative process. Because there is some evidence that the same areas of the brain involved in perception are also active in thought, it is likely that neuron networks situated in the dense local clusters of these brain areas are involved in the process. This aligns well with the principle of functional integration in the brain and efficiently uses cortical areas already specialized in taking in perceptual data to later mentally represent them again. It also leaves us free to say that there is a bigger picture than raw perceptual mimesis involved in cognition.

The aspect of prediction, specifically of using a situated concept, also resonates well with other approaches. Sanford speaks of proxy-situatedness[1] and places more emphasis on situational reasoning, the gist of the approach points in the same direction: situations are learned and stored, then re-activated in order to reason and make inferences when encountering the same situation again, or simply thinking about a situation (Sanford, 2008, pp. 183–184). He then considers the impact that the theory of embodied cognition has on such a process.

Symbols come into the equation by virtue of language and linguistic thought and reasoning, especially when evoking situations in thought or through speech or text. The question is, what exactly does it mean when they say that understanding requires use of the same systems as action and perception? Sanford entertains the notion that this can be interpreted as an absolute, making some kind of bodily motions or actions necessary. His examples are simulating writing movements with one hand to determine the dominant hand, or reciting the alphabet. In other cases a detailed mental simulation of walking through a house, or trimming a hedge may suffice instead of actual action. The third and weakest possibility is that some evocation of situational knowledge brings with it the

[1] Note that for examples of proxy-situations, Sanford refers to cognition only in the sense of situated cognition as defined by Wilson (2002), which is “cognition that takes place in the context of task-relevant inputs and outputs” (Wilson, 2002, p. 626).

activation of cortical motor and perception areas but without the need for a simulation or any action (Sanford, 2008, pp. 185–189). He concludes that at present, while some minimal examples can be constructed for all three levels of embodiment there is insufficient evidence to draw conclusions about whether any of these levels, or direct simulation, are actually necessary for cognitive processing, and that embodiment as a whole is in need of further research and clarification (Sanford, 2008).

The burning question we must ask of both of these approaches is: why do we need embodiment to apply to everything? The implicit assumption in the definition of embodied symbols is that symbols are somehow meaningless if they cannot be 'connected' to a physical object in some way. Read weakly, the definition is trivial, as of course the same perceptual systems used to read, hear and speak are responsible for both understanding and forming language: the brain does it, and the brain does everything, so by definition a process of the brain happens in the brain. Read strongly, the definition seems impossible: The brain has no way of physically connecting concepts to objects, only indirect sensory signals and actions intended toward perceived objects, but those actions are not concepts and they are no more connected to an object than the air is to a tree simply because they are in contact. Many of our concepts indeed do not have a physical object to connect to; love, destiny, never, always, tomorrow, mathematics. Just considering concepts of time, there is an entire array of conceptual items and corresponding linguistic symbols that have no physical equivalent. "An hour" is a perfectly operational mental concept we can calculate with, communicate to others and organize our lives around. There is no object out there in the world that our concept of an hour could possibly be attached to, nothing we can observe. Observing the face of a clock is not a counter-argument; what is observed is the movement of the clock itself, while an hour is our definition of how long it took to complete a certain amount of movement. We could change our definition of an hour, and our observation of the clock would be unaffected. Change the working of a clock, and our definition of an hour would stay the same, and we would call the clock defective when comparing it to another measuring device. Nevertheless, the assumption is in the definition proposed by de Vega et al. (de Vega, Glenberg, & Graesser, 2008) as well as in Barsalou's (1999, 2009) attempt to connect all symbols with a strict empiricism and also implicitly in Sanford's (2008) ideas of embodiment.

Unfortunately, this does not solve the Chinese room argument. Searle's Chinese scribbling fulfilled the requirements of de Vega et al.'s definition already: he perceived the symbols, then perceived the English instructions and used those same perceptual systems to

reproduce the symbols he was shown and fed them back. Searle's Chinese was embodied. The thought experiment is in actuality an exact description of how the brain works and forms conclusions. Compare it to the principles of predictive coding: input, Chinese symbols, comes in. The brain, Searle, has no idea what caused these symbols. It cannot unmix the causes of someone standing outside, or of what the symbols are supposed to tell him. He would not at this point even recognize them as symbols. But the way neurons are connected in the brain give rise to rules: things experienced before are used as priors for things experienced in the future. He learns to recognize the symbols and associate them to the English instructions, forming conclusions about what to represent. Predictive models of the writing being symbolic become recognition models of certain symbols corresponding to others. Searle would be able to learn statistical covariation of certain symbols and he would become quite the expert at recognizing and drawing any combination of them. A graphical representation using predictive model theory would look as follows in Figure 4. Arrows leading from the right to left represent input being added to the current representation via forward connections, encountering predictive model n, while arrows leading from left to right represent the flow of predictions via backwards connections which lead to new predictions and a new predictive model. Errors are represented underneath the input flow, and activation of background knowledge above the input flow.

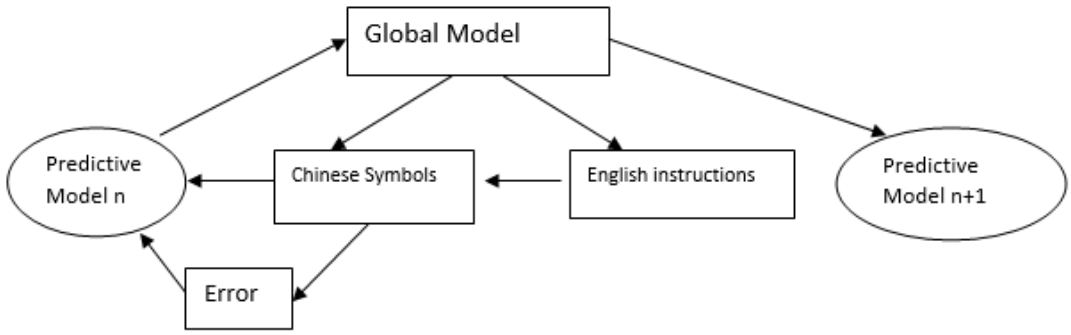

Figure 4. Predictive model n becoming predictive model n+1 through error feedback and input leading to the formation of a new prediction

When Searle is given the Chinese symbols to read, he encounters errors, as he does not understand the Chinese. Some error will be able to be resolved through previous knowledge of his global model. The English instructions are then added, which can be integrated without errors to Searle's global model. Together with the input of the instructions and the remaining error, he forms a new predictive model, n+1, which will now

be used by him to go and form his output of Chinese symbols to the person outside the room – in perfect Chinese.

His causal understanding of Chinese in the sense of what the language is meant to communicate remains poor of course, but his understanding of what the symbols look like, and how to reproduce them indeed becomes as good as any native Chinese speaker's. This is the point we need to take away from this thought experiment. The Chinese room does not prove that symbols need to be connected to the world. On the contrary it proves that we can operate with symbols just fine without needing to attribute a meaning to them. Of course we want to, and our intuitive feeling is that symbol systems are supposed to mean things. The world cannot necessarily solve this problem, because this would only be helpful if Searle both knows exactly what the world is like, and in turn how this relates to the symbols. The difficulty lies in the fact that arbitrariness in this case can mean multiple different things. Symbol theorists who say that symbols are amodal and arbitrary are correct in assuming some levels of amodality and arbitrariness, because there must be a mechanism for us to creatively combine and create meanings from symbols which are not necessarily tied to their original modality or meaning, and Barsalou is correct that we have to account for the original modality of a learned symbol, and that there cannot be inherent arbitrariness in learning a new symbol from the learners perspective, as there must be a form meaning to learn to associate to the symbol in the first place. Of course, if we are taught a new word, then the particular sounds of its spoken form are an integral part of us having learned it, and simply reproducing this would not lead to us suddenly knowing its written form, and vice versa. This does not mean however that the meaning of the symbol we have learned is not still semantically arbitrary and amodal – not all words are onomatopoeias, and learning the verbal form of a word does not mean that what the word means, its actual semantic meaning, is a sound or has anything to do with sound. Barsalou, embodiment theorists and schema theorists are concerned about what connects our symbols to the world, but I believe the fundamental disagreements stem from the fact that the relationship has more stages. *The problem is not only about what connects symbols to the world, it is about what connects the world to our knowledge of it, what connects our knowledge of the world to symbols, and how these same symbols then connect to another individual's knowledge of the world, and finally their knowledge of the world to the world itself*. The solution to the symbol grounding problem is that we do not ground meaning in a strict relationship between mental processes and the world, but between two mental processes – our intentionality towards the world and our internal perception of what the

world is like, the global model. Whether or not the world corresponds to this, is another matter entirely (and I will discuss some pitfalls behind these notions further in section 3.2). Predictive model theory can begin to approach this complex relationship by taking into account individual's global models, and seeing the interpretation and relation of symbols to the world as an active process of predictive models which can change as needed to silence errors.

In fairness to Searle, it was not his intention to prove the necessity of embodied symbols or to argue for an embodiment theory at all. His argument was intended to show that following rules to manipulate symbols was not a sufficient condition for thinking, because following the rules contains no intentional moment to connect concepts to symbols and thus to understand them (Searle, 1980). Intentionality in Searle's sense is also causality, the ability to match mental states to perceived states. This is the crux of the argument I would also make: a program by definition consists of forward connections and recognition models, but it is incapable of forming predictive models. When the program encounters an error, it shuts down. Programmers must always include foreseeable errors within the recognition models of the program. They cannot program inference. Seen from another level, the advantage of the human is to recognize symbols for what they are: perceptual objects that symbolically represent something else. The human brain can do this by learning what the symbols are supposed to look or sound like and by learning what the "something else" is supposed to be then recognizing the symbols as a representation of a representation. It is not a connection to the world which enables this understanding, but a matching of a predictive model about what a symbol is to a predictive model of what the symbol represents and then smoothing out any remaining errors. The error feedback and correction loop is what sets us apart from the machine algorithm, and allows a freedom from the original modality and meaning. Forward connections and error signals allow us to recognize that our understanding of the world, and our learned understanding of the incoming symbols is at odds, and to correct it. All that happens in the brain, or in the Chinese room as it were. Sometimes, neither the meaning nor indeed the symbol are really from the world.

In this chapter I have argued for a separation of processing hierarchies between recognizing symbols and attributing linguistic meaning to them. I have also argued that the symbol grounding problem as discussed in the literature can leads us to conclude that it is possible to learn entire symbol systems without knowing what language, or specifically what system

of intended meaning, they represent. All of these processes can be analysed and explained by predictive model theory.

We have answered research question number 1: what knowledge of the brain can we utilise to describe and understand reading processes and what is predictive model theory? The knowledge that I have presented is the knowledge of neurons and neuronal firing which show us how information travels around the brain. I have presented predictive model theory, and I have begun to show how this process can be seen underlying basic principles of perception and also language, discussing it alongside examples from Barsalou (1999, 2009) and Searle (1980) in this section. We form predictive models against which new inputs are continually compared, and which are ultimately integrated with and compared to our global model. If what we take in through our senses corresponds to already active predictions from the global model or predictive model, forward connections from our senses are silenced and no further processing occurs. If they do not match, errors are propagated, and new predictions formed in cycles to attempt to suppress the errors. With this knowledge, we may continue on to the next research question. In the following and throughout Chapter 2, I will discuss the naturally following question of how meaning then can be attributed to learned symbols and how our knowledge of these interacts to answer research question 2: how can predictive model theory help us to understand the systems underlying language? I will begin in the next section, 2.1, by going back to the issue of symbol grounding and reference to the world and showing how this problem appears even when attempting to relate symbols which are contextualized and modal to the world by looking in more detail at Barsalou's perceptual symbol theory (1999).

Chapter 2: Predictive Models in Linguistics

2.1 From Symbols to Language

In this Chapter I will turn to the next research question as I have laid them out in section 1.1, which is: how can predictive model theory help us to understand the systems underlying language? To begin with, in this section, I will address some of the difficulties in language and our theoretical explanations of language systems and difficulties which require a new approach. I will argue for a theoretical divide between symbols as worldly objects, our internalised knowledge of these symbols and their associated semantic meanings and linguistic systems, by turning to a theory of linguistic symbols from psychology. I will argue for this separation based on the difficulty of knowledge and association of such knowledge to symbols, and limitations of our worldviews and perceptions. The following sections of this chapter will address other aspects relevant to an answer to our research question, with a final response and summary at the end of Section 2.4.

Symbols and language are generally treated by embodiment theorists as one system. By doing so, they inherit the problem of the Chinese room described in Section 1.5. A good place to make this obvious is in the perceptual symbol theory of Barsalou (Barsalou, 2003; 2009). I did not choose Barsalou here because I believe that he is wrong. In fact I strongly agree with a majority of his ideas and they will form a part of the later theory described in this work. The missing divide between symbol and meaning must be eliminated from the theory however. We will start at the beginning.

Wu and Barsalou (cf. Wu & Barsalou, 2009) ran a series of experiments on background knowledge and specifically knowledge of actions associated with words by taking groups of participants, providing them with a noun and asking them produce as many properties of the noun as possible, e.g. a noun such as “sky” might lead to the properties “blue”, “clouds”, “air”, “large” etc. These are known as property generation experiments. The researchers were interested in whether participants would produce properties which were closely associated to common situations these objects are encountered in, because they may be mentally “simulating” the contextual situation in which one encounters the object. Evidence from such a property generation experiment suggests that perceptual simulation

may occur when combining concepts and attempting to generate properties for objects which they only possess under specific circumstances, such as occlusion (Wu & Barsalou, 2009, p. 174). The researchers predicted that participants generating properties for nouns such as “lawn” in isolation should produce different results than when producing properties for combinations such as “rolled-up lawn” (because of the density of grass roots, lawns can be grown on sheets of turf which become thick enough to be rolled up into cylinders and sold, which allows for the bottom layer of turf as well as the grass roots to be visible), because the conceptual occlusion would be cognitively simulated, thus occluding properties like “roots” or “dirt” (Wu & Barsalou, 2009). Results across three experiments seemed to support this view, with participants generating occluded properties far less frequently than unoccluded properties. When occluding surfaces were removed from target nouns, internal, previously occluded properties inversely became far more frequent. Participants also produced high numbers of properties describing background situations for nouns rather than direct properties of the objects. The conclusion was that simulation is an integral part of property generation (Wu & Barsalou, 2009). We can describe this phenomenon in terms of predictive model theory as well, whereby the simulation of an action is an instantiation of a predictive model activated by a consideration of specific entities. Since predictions are based upon priors formed through past experience, they include likely physical orientation and background information, including information about internal properties which would be occluded. As discussed before, having detailed predictive models about such situations is important because perception itself would not be able to resolve the occlusion. Simulation may form one method of using predictive models, especially within visual perception. There is however a looming issue in perceptual symbol theory, with implications for predictive model theory.

The jump from describing pure perception to linguistic systems is rather subtle in the literature reviewed so far. Much of the predictive coding literature shies away from language altogether; any and all evidence is based squarely on perception and even then the focus is predominantly on vision (with the exception of one example by Friston which will be discussed in a later section). In Barsalou we see first steps of nouns and linguistic content being related to perceptual symbols, which Barsalou defines as schematic memories of a particular object or situation, but only in a specific way. The study on occlusion and simulation described above appears to make a very fundamental supposition that is also apparent in much of Barsalou’s other work: nouns purely refer to perceived objects. This is a thesis going back a very long way, which is riddled with difficulties.

Philosophy has grappled with it for a very long time, coming to a head with Frege's distinction between sense and reference (Frege, 1948). Barsalou refers to this, directly noting that perceptual symbols fulfil the requirement of establishing sense and reference and citing Frege (Barsalou, 1999, p. 597). This is not an advantage for the theory unfortunately.

The impetus for Frege's distinction comes from an apparent mismatch between identity statements and referents. Suppose that we had two names (linguistic symbols) for one object, but never knew this before. If we were to discover that they are indeed the very same object, we would declare them to be the same thing. Statements of the form 'a = a,' and 'b = b' are tautological and impart no knowledge. The new statement 'a = b' is meant to be a discovery, such as discovering that the morning star and the evening star are in fact not stars, nor two objects, but the planet Venus visible in the sky. This poses a problem for us if we assume that names simply stand for their referents in a fixed way. Saying 'the morning star = the evening star' would amount to a tautology we should already have known, since they refer to the same object, or it becomes a statement only between the words themselves. Names and words however are arbitrary, so again both options amount to the same (Frege, 1948, pp. 209–210). Both options put us in the absurd situation of wanting to say we learned something, but being forced to say that since the referent was fixed by the name, we already knew (even though we obviously didn't). Frege's highly controversial solution was to add another aspect to the relationship:

> It is natural, now, to think of there being connected with a sign (name, combination of words, letter), besides that to which the sign refers, which may be called the referent of the sign, also what I would like to call the sense of the sign, wherein the mode of presentation is contained. […] The referent of "evening star" would be the same as that of "morning star," but not the sense. (Frege, 1948, p. 210)

While this solves the immediate problem of simply endlessly equating arbitrary signs with one another, it opens the door for an even larger problem. Though we may know the referent of the name and generally speaking at least one sense of it, real objects will have many possible senses and names which we cannot realistically all know. To confound things further, Frege does not believe that natural language always follows these rules. Even worse, some names or designations may have a sense but no clear referent, if they refer to a theoretical entity such as 'the smallest number' or 'the longest line.' He concludes: "In grasping a sense, one is not certainly assured of a referent" (Frege, 1948).

Philosophers have since added many things to Frege's account, sparking the great linguistic turn in philosophy and producing vast volumes of work. Kripke later argued that in giving

such a definition to the process of referring, we must also implicitly assume a familiarity with an object: “Every time we determine a referent, we are introspectively acquainted with how the referent is determined, and that is the corresponding sense” (Kripke, 2008, p. 199).

If we think back to Wu and Barsalou (2009) and the experiment on conceptual simulation above, and the assumption that nouns refer to perceptual symbols, we can more clearly relate perceptual symbols to Frege’s (1948) sense. The mode of presentation when taking in certain objects in the world would be perceptual, so akin to the sense of the morning star being the particularly bright object visible in the morning, the sense of “lawn” and “rolled-up lawn” would be the experiences of lawns and rolled-up lawns. The referents would then be actual lawns and rolled-up lawns. This is already problematic. Clearly there is not a single referent that all human beings could be referring to, although the experiment clearly expected there to be a parallel. Let us sympathetically say that it is because the researchers are running on the assumption that there is a singular material world out there, with material objects. Therefore, individual sense data would in fact reflect more or less the same objects. Since there are types of objects in the world, such as lawns, which share many basic properties, the observed properties would be the same wherever or whenever they may be observed, thus leading to more or less the same perceptual symbols about them in observers. This is a reasonable assumption, which I share. What happens next is what poses all the problems. As stated, the method of observation would grant a sense of each noun in Frege’s (1948) definition and it would fulfil Kripke’s (2008) elaboration of Frege: by identifying that we are talking about a referent, the lawn, we also identify that we know about lawns by having seen them, touched them and so forth, thus getting the sense of lawn. This can clearly be inferred from Barsalou’s treatment of the relationship between his perceptual symbols and referents:

> Once a perceptual state arises, a subset of it is extracted via selective attention and stored permanently in long-term memory. On later retrievals, this perceptual memory can function symbolically, standing for referents in the world, and entering into symbol manipulation. (Lawrence W. Barsalou, 1999, p. 578)

This basic assumption about perceptual symbols having direct referents is also in the description of Wu and Barsalou’s account of simulations representing the referents of concepts and thus linking a concept such as “house” with its perceptual and introspective experience (Wu & Barsalou, 2009).

Prima facie, this seems fine. It is in fact more fine-grained than what Frege had originally talked about, as Barsalou and colleagues do not make the obvious mistake of saying that

each concept, as they put it, refers to one singular object. Rather it is a concept of houses, formed by an amalgam of repeated experiences. All of these examples work quite well because we are talking, as some of us often jokingly observe in ordinary language philosophy, about 'medium-sized dry goods.' Lawns and houses are safe examples of easily delineated things, so it is also quite easy to talk about them, conceptualize them, and most importantly, call them lawns or houses. This does not however answer the question: why do their perceptual symbols specifically contain the words lawn and house? How does this account for synonymy? In order to complete the jump from the perceptual data to a linguistic sign, in accordance with this theory of conceptualization, there needs to be some systematic relation between the sign and the referent.

Barsalou attempts the stretch by making words themselves into perceptual symbols:

> In humans, linguistic symbols develop together with their associated perceptual symbols. Like a perceptual symbol, a linguistic symbol is a schematic memory of a perceived event, where the perceived event is a spoken or a written word. A linguistic symbol is not an amodal symbol, nor does an amodal symbol ever develop in conjunction with it. Instead, a linguistic symbol develops just like a perceptual symbol. As selective attention focuses on spoken and written words, schematic memories extracted from perceptual states become integrated into simulators that later produce simulations of these words in recognition, imagination, and production.
>
> As simulators for words develop in memory, they become associated with simulators for the entities and events to which they refer. (Barsalou, 1999, p. 592)

Such an explanation holds and is backed up by neuroscientific evidence. Pulvermüller for example speaks of "action-perception networks" which link motor and auditory neuron assemblies together to represent and process specific spoken word forms (Pulvermüller, 2008). These networks which link word forms with the neurological networks necessary to identify them from auditory input, and with motor systems necessary to produce them in spoken or written form, are backed up quite well; see (Pulvermüller, 2008; Pulvermüller & Preissl, 1991; Pulvermüller, Shtyrov, & Ilmoniemi, 2003). All that a network such as this describes however, is how to recognize and produce the word form in question. When it comes to the actual association with this word form and semantic meaning, Pulvermüller resorts back to classic Hebbian learning, with a conservative suggestion that: "Hearing the word 'crocodile' frequently together with certain visual perceptions may lead to strengthening of connections between the activated visual and language-related neurons" (Pulvermüller, 2008). Action words similarly may be learned in infancy by the child performing an action then hearing an action word said by a caretaker (Pulvermüller, 2008), see also (Tomasello & Kruger, 1992). The conclusion would be that we learn to associate words with meanings by statistical co-occurrence. While it is plausible that word forms, i.e.

knowing how words sound, look, how to write them etc. do resemble perceptual symbols, there is not a clear perceptual link between a specific word form and a meaning. When Barsalou (1999) says above that simulators for words become associated to the entities and events to which they refer, this suggests that the relationship of sign and referent is set and just needs to be discovered. This does not follow from the theory of learning perceptual symbols however, which can only really say that signs and referents are 'usually' experienced together.

Returning to the morning star and evening star dilemma, we can see that perceptual symbols fall into the same trap as Frege did. A keen observer of the sky with no other knowledge of astronomy may well develop two designations for bright objects, based on the context of when they were observed. This observer chooses to call one "the morning star" and one "the evening star," while associating these designations with perceptual impressions of the two. It would be quite natural to suppose that the observer has formed two perceptual symbols, one standing for the object seen in the morning, and a second standing for the object seen in the evening. The two linguistic signs chosen have become associated with their respective perceptual symbol. Now our observer learns that they are one and the same. If we assume that perceptual symbols can establish only reference, clearly we are again at a tautology. Our observer should have known they were the same, since both perceptual symbols referred to the same object. That is obviously not correct, so perceptual symbols also need a sense in order to give meaning to the statement that the referents are the same. The result would perhaps be something like "The object seen in the morning sky is the same object seen in the night sky." In predictive model theory this is not really much of a problem. Each observation of the object is a minimal predictive model, wherein one object, the bright observable "star" in the sky is observed in context of a particular location and time. The morning star is a bright object observed in a particular spot in the sky in the morning hours, the evening star is a bright object observable in a particular spot in the sky in the evening hours. Upon learning that these two symbols refer to the same object, we may receive an error, but then we will be able to integrate it into our global model as in Figure 5.

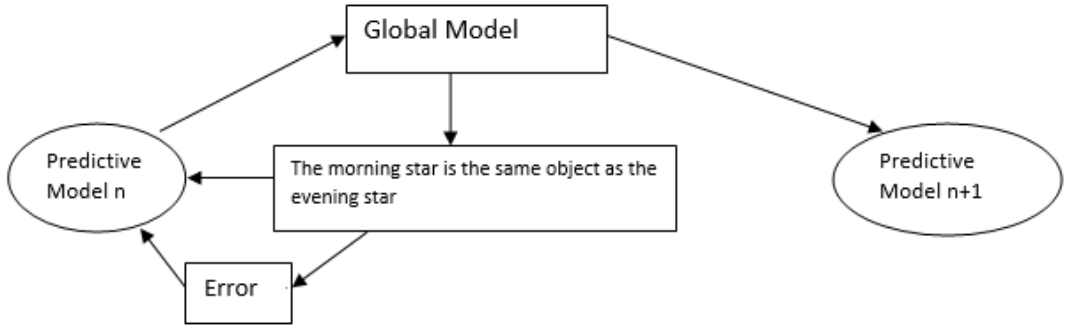

Figure 5. Learning that the morning star and evening star are the same object leads to new predictive model n+1

This integration works, as these minimal predictive models can overlap without any conflicts. While the object is the same, the spacial and temporal locations are unique and so they may form a new predictive model in which the object is viewed in different places at different times. This new predictive model can now integrate with the global model, and further our understanding of the actual observed object, exactly as Frege (1948) intended. In fact, I believe that what Kripke was appealing to in the wording of "we are introspectively acquainted with how the referent is determined" (Kripke, 2008, p. 199), is this global model integration. We cannot introspectively appeal to real objects, as those are not in our heads. What we can appeal to is the context of how we have come to associate certain symbols or perceptual impressions to specific meanings, and upon this introspection we can see that it is entirely possible to know that the morning star and evening star refer to the same object, but also maintain that they have semantically different and fruitful meanings, to describe the object seen in the morning and evening and its relative location in the sky respectively. We also have not had to alter our perceptual interpretations of what this object looked like at all, as the operation was purely at the level of interpretation and semantic meaning of these descriptions. Now we no longer have a tautology, but perceptual symbol theory has a difficulty.

While word meanings and descriptions can change without modifying the referent, perceptual symbols seem to have a complication: "Because perceptual symbols bear structural relations to their referents, structural changes in a symbol imply structural changes in its referent, at least under many conditions" (Barsalou, 1999). This statement, given the previous adherence to Frege and a stark realism about the world, coupled with a continuous insistence that symbols always refer to objects in the world, is extremely

puzzling. Even more puzzling then that somewhat earlier in the same text, a much weaker definition is given, whereby perceptual symbols can refer to a wide range of things:

> As we shall see later, the *designation* of a perceptual symbol determines whether it represents a specific individual or a kind – the *resemblance* of a symbol to its referent is not critical […]. Suffice it to say for now that the same perceptual symbol can represent a variety of referents, depending on how causal and contextual factors link it to referents in different contexts […]. Across different pragmatic contexts, a schematic drawing of a generic skyscraper could stand for the Empire State Building, for skyscrapers in general, or for clothing made in New York City. (Barsalou, 1999, p. 584)

This is not only an ontologically confusing combination of things but only actually lists one identifiable world object out of three. “Skyscrapers in general” is a theoretical entity or a concept that really has no necessary corresponding object, not any more than “unicorns in general.” Similarly, “clothing made in New York City” is a category, also a theoretical entity, which includes past clothing, present clothing and any future clothing manufactured in that particular city. If anything, the observation of a category of objects sharing a particular quality, in this case being made in a particular place, should be another perceptual symbol. This is in fact what I would argue.

The only possible conclusion that could make sense of these disparate things is that perceptual symbols don’t directly refer to real objects but to our predictive models about objects, as well as to our network of predictive models or perceptual symbols. They can refer, in a specific context, when a referent is readily available to provide the necessary familiarity as according to Kripke (Kripke, 1981; 2008). In practice, this means when there is an object providing the empirical input to which the predictive model can be equated, or for instance another symbol which then in turn refers to the object, forming a chain of reference. We can reformulate the idea of perceptual symbols as referring to objects to instead be priors, which refer to entities postulated to exist by a predictive model. Repeat perceptual experiences which caused driving sensory signals have caused an observer to learn recognition models for these inputs, which have been habitually paired up against a predictive model which posits there being a cause for the signal, and that the cause is the existence of an object. Because such a predictive model was very good at explaining the driving signal, it has been accepted in the past. The explanation behind the perceptual phenomenon now causes the observer to have to undo some of that association and form a new predictive model of there being one object causing different sensory signals at different times and to correlate it to both learned linguistic symbols. This is possible and will be possible every time an error is identified precisely because the linguistic symbols have no necessary structural ties to any given concept. If there is not an object but only an

abstract, or if there is a chain of reference i.e. talking about a word in a dictionary which in turn refers, then we can resolve these cases by either referring to our abstract knowledge, or by resolving our predictive model to be satisfied by the referent chain instead.

As such, perceptual symbol theory gives us a very powerful tool for evaluating both how symbols and meanings are represented at the word form and meaning level, while simulation ties in well with the predictive model level of understanding. It also helps us to refine our answer to the symbol grounding problem from section 1.5. If it were true that our internal knowledge must be grounded on a relationship between our consciousness and the world, we would not be able to have these distinct perceptual symbols of the morning and evening star, but would ultimately have to abandon them for a single representation of the singular object. Instead, we are quite capable of having all three – we can conceive of the morning star, the evening star and at the same time the single planet which it is we are observing at the specific times of morning and evening. To get to this point we have had to abandon several aspects of the theory.

Both the word forms and word meanings represent kinds of perceptual symbols as Barsalou (Barsalou, 2003; 2009) envisions it, and it must certainly be the case that the modality of each kind is linked to the way it is stored, because it is naturally linked to the way they are learned. Spoken word forms are taken in by the auditory system and thus would be anticipated by localized predictive models within the auditory cortex for example, leading to a modal localization of predictions about this word form. Therefore, a spoken word form would indeed become a perceptual symbol stored within the areas of the brain specialized for the intake of its modality. Once we learn to also articulate this word form ourselves, other parts of the brain must be recruited in order to properly carry out the required functions, leading to another perceptual symbol located in articulatory and motor control areas, and these are the kind of assemblies that Pulvermueller (2008) speaks of. We must be very careful in our classification of these however: knowing how to say a word does not constitute knowing what that word means and even experienced speakers can find themselves surprised by unusual word forms. For example, a good rule of thumb for many words is that the negative of an adjective has the prefix ‘in’ or ‘im’ such as inflexible, impossible, intractable – except of course for inflammable, which means flammable.

Some studies, including by Pulvermueller (2008), claim that certain phenomena such as the action-sentence compatibility effect, or ACE, is evidence that because areas of the motor cortex become active when listening to and comprehending sentences including action

words, meaning is stored in motor areas. This is the core of embodiment theory. The ACE effect is however not robust and does not occur during processing of metaphors or overtly fictional situations, but this makes sense if we take my distinction of the different knowledge networks as argued in this chapter into account. The written word form of "ring" for example is tied to the visual or haptic modality for written letters or braille letters respectively, but the meanings of this one word cover a wide spectrum of modal information. The issue begins in even deciding which possible meaning of ring to take: physical object, action, sound produced, geometric shape? What modality is a geometric shape? It is clear that in some contexts the background knowledge of auditory perception will be needed to resolve reference to say the ringing of a phone or the ringing of a hammer upon an anvil, but not for a ring drawn on paper with a pen, or a ring of objects around another one. Some meanings are regional as well, as to "ring someone" is commonly understand to mean calling on the phone in the UK, but not widely used in North America for example. In case of metaphor and fiction one might make the same argument. In the idiom "to kick the bucket" we do not mean a literal kick, and the more entrenched such phrases are the less likely any literal information would be needed to resolve its meaning. Similarly, if one is reading about a fictional action in a piece of literature which the reader is not able to do nor has any information on how it is done, it would not be possible to activate the motor cortex to simulate it. While it cannot give satisfactory evidence for action word meanings being stored in the motor cortex as a general rule, The ACE effect does help give further credence to the fact that any given word must consist of an amalgamation of many differently distributed perceptual symbols, of which the recognizable word forms in written or spoken form are but a part.

This solution can be compared to a similar, documented case of error activity in vision. Egner, Monti and Summerfield tested whether activation in the Fusiform Face Area (FFA), an area of the brain well associated with recognizing faces, was better explained by predictive coding or classic feature detection theories, which postulate a single process of neurons recognizing and encoding input. To accomplish this, subjects were shown pictures of faces or houses inside coloured frames. An orthogonal control task was for the subjects to press a button if a stimulus was presented upside down. The images were presented inside coloured frames, with the frame being visible 250 ms before the image. The colour of the frame indicated the probability of the picture being a face or a house, ranging from high probability of faces through equal probability of face or house and finally a low probability of the pictures being faces. Subjects were verbally informed of this likelihood beforehand,

but not of its significance (Egner, Monti, & Summerfield, 2010). Using predictive coding, the researchers predicted that activation in the FFA should be similar under high face expectation whether the subject sees a face or a house, and maximally different under low face expectation (Egner et al., 2010). The results confirmed that the activation patterns of subjects' FFA, measured by BOLD (blood oxygen level dependent) fMRI did indeed show minimal difference between face and house stimuli when the subject had high face expectation, and maximal difference under low face expectation (Egner et al., 2010). In other words, there was a very small amount of activity in the FFA when a subject expected to see a face, then saw either a face or a house. In contrast, there was a large amount of FFA activity when the subject did not expect to see a face but was shown one, but only a small activity when the subject did not expect to see a face and saw a house. The explanation for this phenomenon is that error signals and predictions are localized and occur in different regions for face and house surprise (Egner et al., 2010). While expecting a face, a predictive model is already active, resulting in mild FFA activity. Seeing a face in this state elicits no error, because the driving sensory signal is already matched, causing no further activation. Seeing a house in this state also causes no further FFA activation, because the FFA does not deal with recognizing houses; the error signal is elsewhere in the brain. Under low face expectation, the active predictive model does not contain a face. Seeing a face elicits an error in the FFA, which is specialized for signals resembling its recognition models, causing it to have to process the error and match it with a new prediction, resulting in high activity. Seeing a house in this state still elicits no further FFA activity, as the error is still elsewhere in the brain.

This particular kind of functional segregation is what I believe to be responsible for the fact that errors in word forms can have minimal effect on semantic understanding: errors elicited by a word form may well be resolved within the assembly responsible for recognizing and producing the word, without eliciting errors in the conceptual areas responsible for assigning meaning, and vice versa as we saw above with the morning and evening star example. This kind of resolution has also been suggested in auditory speech perception, with hearers predicting a word rendered inaudible by noise or poor enunciation (Garrod, Gambi, & Pickering, 2014), which is an analogue of visual occlusion. The conclusion to all of this must be: *linguistic symbols are a complex system unto itself, learned by the brain as certain visual, auditory or tactile impressions which can be recognized and reproduced, and then associated to real world objects if possible, or otherwise to other mental concepts, perceptual symbols, or chains of symbols which eventually represent a*

world object. As we have learned from Searle's Chinese room example in Section 1.5, we may learn to manipulate symbols without knowing any meaning attached to them. Together with the meaning we attribute to the symbol, a predictive model is built and tested against the global model, but the exact contents of our representations at the symbol level and meaning level can both be negotiated online, within the predictive model. We can represent this as follows in Figure 6:

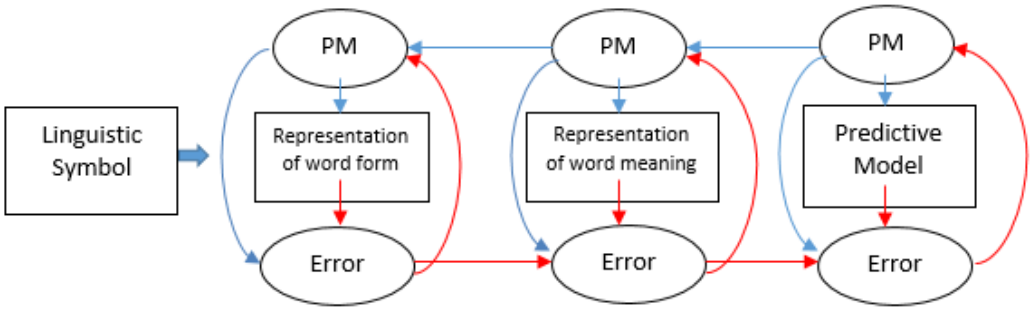

Figure 6. The prediction and error cycle leading from perception of symbols to predictive models

In this Section I have argued that we must consider a theoretical divide between objects and words so as not to run headlong into issues of reference as faced by Frege (1948) and other theories of language. Instead, I have suggested that symbolic systems can be learned independently of their meaning, and are first associated to our mental concepts of the world. They then form complex networks of predictive models with other symbols as referents, or chains of symbols which refer to our representations of objects. This is always resolved at the stage of matching prediction to incoming signal, and can thus be contextually refined to activate the appropriate knowledge if available. Over time and usage, our associations from symbol to world object will become more refined, and because we of course live in the world while learning both about its objects and the human symbols used to represent them, these processes will overlap. For an explanation of the process however, we must didactically separate them. In order to now integrate word forms into the understanding process, we must consider in more detail how they relate to the next section of the model, word meanings. To discuss these in more detail I will refer to cognitive grammar as envisioned by Langacker (1986).

2.2 From Words to Meanings

In the previous section I discussed the learning of symbols and symbolic systems and the separation we must make in theory between them and the meanings attributed to symbols. The aim of this section will be to discuss the association between meanings of words and symbols in more detail and to relate them to our predictive model theory framework. I will also argue that a predictive model analysis is necessary for resolving word meanings in utterances which rely on contextual factors. To do so, we will turn to cognitive grammar, and a fascinating field of study surrounding the phenomenon of pragmatic normalisation.

Some meanings can be represented by words but also by direct sensory input. As a trivial example, our concept of a chair can be triggered by hearing the word chair but also by simply seeing a chair, or a picture of a chair. Multiple layers of representation are possible. The representation itself of our prototype triggered by simply encountering the word might be different in some respects to our visual representation of actually seeing a chair, but the brain still uses a predictive recognition model in order to recognize and represent both as chairs. The difference is that using the word forms a dual representation of the symbol together with the representation of the assigned meaning, while seeing a chair matches the predictive model directly to the sensory input. The sensory input of the chair is substituted by using the word instead, together with any given amount of description. Speaking about a mahogany chair with a straight back and a brown plush seat conjures a reasonably specific mental image, actually seeing such a chair perhaps even more so, but just saying chair also suffices for some kind of mental representation. The requirements of a predictive model in finding and building an acceptable match to such input are that an individual has some existing knowledge about them, or at least enough contextual knowledge to form a representation. One very useful way of investigating how meaning might be formed from such linguistic input comes from cognitive grammar.

As a better contrast to the cognitive theories discussed so far, we will begin with Langacker's ideas of how language and linguistic signs might be learned from the viewpoint of cognitive grammar. Much of it aligns quite well with perceptual symbol theory:

> Learning a language consists in mastering a large inventory of patterns of activity. These patterns are of various sorts: motor, perceptual, conceptual, interactive, or any combination thereof. A thoroughly mastered pattern is what I call a unit, which is thus a chunk of linguistic expertise. (Langacker, 1986, p. 628)

These units reside in recurring patterns of activity, with each linguistic unit being constituted by the same processing activity required for its production. As linguistic expressions are encountered and produced, units become abstracted from them and stored through Hebbian learning (Langacker, 2009). Here too we can clearly see that the idea of individual words or units being selectively stored from experience echoes perceptual symbols. The abstracted linguistic units can then be activated during new experiences of linguistic expressions to identify expressions and categorize the recognized units (Langacker, 2009), akin to a predictive model. As a result, when something is perceived, in Langacker's example an object which may be called either a cup or a mug, existing units associated with its features will be activated with the units competing for activation, until the one whose features most closely match the input "wins" (Langacker, 2009). This can be translated easily into the terminology of predictive coding: top-down predictions become active to attempt to explain away and categorize what is perceived, using existing perceptual schemata as priors, with error signals from the input helping to select which predictive model is active.

The process of learning word sequences follows the same mechanic as for single units, with word sequences being abstracted from larger expressions, and certain words being associated through frequent use (Langacker, 2009). In this sense Langacker constructs language as a neural network, with complex expressions amounting to more complex network structure, but without specific localized information:

> A key point is that complex units overlap, consisting in partially shared processing activity, rather than being separate and disjoint. They are not stored in separate places, indeed they are not per se stored at all. Rather they occur, consisting in recurring patterns of activity. Ultimately, a unit resides in adjustments made to synaptic connections, permitting the occurrence of patterns of activation similar enough to be functionally equivalent in some respect. (Langacker, 2009, p. 635)

This is essentially a formulation of predictive coding. Rather than existing in some place, linguistic symbols and meanings are given by predictive models, with backwards connections modulating synaptic plasticity to achieve the desired output configuration, and often-used predictions becoming ingrained through Hebbian learning as predictive models for future input of a similar kind. No further supervision or guidance is required. This phenomenon is a direct and important part of theories of computational cognition, which show quite well that neural networks learn to use rules without needing to store the rules themselves (Spitzer, 2008). It has also been a long-standing argument in ordinary language philosophy since Wittgenstein (2006), who argued that there is an epistemic difficulty in discovering if someone else has understood a rule as an abstract quality, since all we can

test is asking the other person to solve specific examples –there is no firm proof of "understanding" of the rule itself (Wittgenstein, 2006: pp. 316-339; see also Holtzman, 2014). In light of what we now believe about the process of learning and neurological association, this view makes a lot of sense. We learn language through usage, as Langacker (1986) suggests, that is through individual examples and we, as learners, suppose that there are rules underlying the language whenever several examples seem to work in the same way. We use these to try to predict how new utterances should be formed. In practice however, there are few such rules which can always be applied without exceptions.

Words themselves may then be considered as a gateway or index for a portion of our stored knowledge (Croft & Cruse, 2004). The meaning of a word is essentially represented by a neural network of meanings and associations, with a definite meaning only ever being decided in an online "construal" which is influenced by a long list of cognitive operations and factors, including attention, salience, but also identity, viewpoint and deixis, boundedness and more (Croft & Cruse, 2004). This view makes it clear that while there are indeed specific neurological networks for the production of word forms in speech and writing, the semantic meaning of the words per se contains far larger knowledge networks. I shall somewhat modify then adopt Langacker's (1986) idea of words acting as index markers for overall knowledge networks. Rather than a word coming to have a specific meaning, it seems indeed right to say that a word comes to be associated with, and thus indexically stands for a field of knowledge structures. I would suggest that these are the kind of structures I discussed in Section 2.1, regarding the modal qualities of the word form, its production, its reference and any associated meanings or other knowledge we might have of it.

Because of the gradual nature of human learning and the almost boundless flexibility of language to act in this way, words become subject to the phenomenon of polysemy. It can be defined as the isolation of different parts of a words' meaning potential for specific construals (interpretations) of the word in context (Croft & Cruse, 2004). As such, words can have bounded sense units which favour a portion of the overall meaning while suppressing another, as in the word "bank." Depending on the context in which it is uttered, bank may stand for an institution dealing with the storage and management of money, for a specific instance of a bank branch located on a high street or elsewhere, for a bounding portion of land next to a river, for a pile of earth or clouds, for a row of similar objects such as machinery or parts, for an institution or specific location for storing other objects such as blood or sperm, in gambling it represents money belonging to the owner, as

a verb it can mean storing or winning money or other objects, to fly an airplane with one wing raised up above the other on a horizontal plane, or to collect or form something into a mass or cause it to do so ("bank - definition of bank in English from the Oxford dictionary," n.d.). This is a very wide array of potential interpretations and it is very clear that several of these possible meanings cannot be held to be the meaning of the word in use at the same time. We would not usually talk about mooring a boat to a financial institution (Croft & Cruse, 2004, p. 109), or of storing our money in a bank of clouds, or say that airplanes were particularly concerned about investing their money. Croft and Cruse argue that sense boundaries, in which a specific meaning from the field of possibilities is picked and bounded out against them, are not a property of the word, but only drawn during the interpretation, that is, when actually being confronted with the word and having to form an interpretation (Croft & Cruse, 2004).

I wholeheartedly agree with this idea and we can use predictive coding to explain it quite plausibly from a neural perspective: through repeated experience word forms become associated to conceptual meaning structures which can become quite large and diverse, owing to a limitation of how many word forms exist compared to the much vaster number of possible concepts. The word form itself has only an arbitrary connection of referential usage or reference chains which associates it to the meaning potentials in the first place and which can change over time or connect it to multiple referents or meanings. When encountering the form, the context of the word form is taken into account and a resulting predictive model formed which attempts to match what the symbol is intended to stand for in that specific utterance. The process of selecting an appropriate predictive model both creates specific error signals, defined by mismatch against the prediction, then silences them through more specific predictions. This very process causes what Croft and Cruse (2004) call antagonistic meanings, and the ultimately bounded construal. The actual neural mechanism shuts down the noise of competing meanings by inhibiting their synaptic pathways and attempts to find the closest predictive match, the best construal, which is then selected as the representation of what is meant. In this way the neural mechanism of recognizing the word quite naturally activates the parts of the knowledge base positively associated to the word form, then eliminates implausible interpretations and arrives at a contextually selected "optimal" interpretation based on the predictions available. This process indeed has nothing to do with the word form itself, and is not contained in the word form. We might argue that a certain meaning is intended by an author of an utterance, and this is usually achieved by placing the words in the correct sequence and

context, but the actual process of arriving at the intended meaning takes place within the brain of the receiver. The word forms could still stand for anything, and are subject to misunderstanding by a receiver.

Of course we do not forget the other possible meanings of a word while deciding on one definite interpretation and this becomes readily apparent in wordplay, such as when I suggested above that one might talk about airplanes investing or managing money, or mooring boats to a bank outlet. It is also possible to construct fictional scenarios where such an interpretation is right: imagine a bank building situated directly by a river, accessible by boat, perhaps with a small quay for mooring a boat to go inside. In such a context we could see two possible interpretations being considered *correct* when saying it was moored on the bank. We can also generate zeugma by saying it is a bank on a bank. These processes are caused by deliberate imprecision being built into the utterance structure, which in turn causes larger errors and more complicated predictive models to compensate, allowing for multiple divergent explanation to satisfy the input. This diverse range of meanings activated against the index of words has great explanatory power but also must be constrained somewhat. When considering what a speaker might understand within the utterance of a word or string of words, we cannot profile the possible meanings against a dictionary listing of word meanings. The dictionary list contains a listing of standard means as codified by an authority – in the above case I cited the Oxford Dictionary website. An actual speaker may not necessarily be aware of all possible facets of a word's meanings, or might even be mistaken about a meaning (although saying mistaken only holds true when accepting the authority as the only correct way to interpret word meaning). Obviously word changes and novel ways to interpret words also play a part in differing interpretations.

We can also see the divide between word form and meaning in the phenomenon of pragmatic normalization. It is a phenomenon in which knowledge and meaning interpretation may override the actual words, or what Sanford and Emmott call local semantic analysis (Sanford & Emmott, 2012), of an utterance. They point to very convincing evidence for this phenomenon: studies found that disordered conjunctive sentences such as "John dressed and had a bath" were normalized when paraphrased by 64% of subjects, leading to more sensible interpretations of the sentences, with only 42% of subjects reporting that they recognized a difference in meaning between the disordered and normalized sentences (Fillenbaum, 1974). Another study suggests that when faced with syntactically difficult passive clauses, such as "The dog was bitten by the man" participants

asked to make judgements about who the thematic agent of a sentence was, tended to make mistakes in correctly assigning agency and to attempt to normalize the sentences to fit more plausible scenarios (Ferreira, 2003). Similarly, the sentence “No head injury is too trivial to be ignored” was shown to be “systematically misconstrued” as in fact meaning the opposite of what the sentence means semantically (Wason & Reich, 1979). The literal meaning would be that all head wounds can be ignored, as even the most trivial is not too trivial. The normalised meaning would be to construe the sentence as meaning that no head wounds should be ignored, i.e. to reform it as ‘No head wound is trivial enough to be ignored’. Sanford and Emmott relate this to their scenario-mapping theory, suggesting that a scenario is activated like a frame and the words mapped into the slots of this scenario, overriding local semantics (Sanford & Emmott, 2012).

Fillenbaum suggests that participants likely based their responses on the assumption that discourse is intended to describe situations sensibly, following customary orders of events and causal relations. They thus assumed that differences between target sentences and normalized sentences reflected not a difference in described events, but a failure by the producer of the target utterance to describe the intended event properly (Fillenbaum, 1974). Ferreira turns to a strategy of “good enough” processing and investigates processing along two parallel lines, heuristic and algorithmic (Ferreira, 2003). The basic assumption of this view is that readers or listeners often engage in shallow processing because they reach a stopping point at which a decision about an utterance is made, even if it is inaccurate and incomplete (Ferreira, 2003, p. 168). A way to categorize how this might work is to propose that there is an algorithmic process which would scan an entire utterance and come out with a “correct” conclusion of the entire utterance meaning, thereby potentially overcoming the normalization and reaching the undesirable conclusion. This process is costly, because it requires a lot of processing, so there may be a second process in place, which we can call a heuristic. In this case the heuristic is labelled the noun-verb-noun or NVN strategy, a mechanism that is supposedly applied by readers or hearers of an utterance in which they make the assumption that within a given utterance there must be a noun which describes the subject on an action, a verb describing this action and another noun describing the object of the action. The heuristic simply assigns these roles to the most likely (read: most often encountered) nouns and verbs in a sentence, defaulting the subject role to the first noun and the object role to the second and forgoes syntactic analysis (Ferreira, 2003). Ferreira’s conclusion of three different experiments involving the switching of thematic roles, passives and cleft structures showed that participants had

difficulty overcoming the assumptions of the NVN strategy and systematically had difficulty with sentences which switched the order of agents and patients (Ferreira, 2003). Nevertheless, the strategy can be overcome, making it likely that both a pure syntactic algorithm and a heuristic such as the NVN strategy were applied, although Ferreira leaves it as an open question how they might interact (Ferreira, 2003).

Predictive Coding can provide a way of explaining the interaction. Heuristic models such as the NVN strategy bear a strong resemblance to entrenched predictive models of often encountered grammatical structure. If this structure is somehow violated, errors occur and are sent onwards in order to be resolved. A predictive model must now be formed to make sense of the utterance somehow, and the hearer must make a choice. In all of the examples above, subjects are confronted with difficult or implausible sentences, about whose meaning the predictive model must make a decision. Rather than one, there are two kinds of errors which occur from such sentences: formal syntactic errors of word order not conforming to the desired meaning, or the desired meaning of the word order not conforming to global knowledge structures. Repairing the first outcome would result in normalization, with the desired meaning overriding errors of syntax but delivering a good semantic fit. Repairing the second outcome would lead to a formally accurate reading of the syntactic structure but with a semantic interpretation that is anomalous, compared to existing knowledge. It is correct that the processing of a sentence, or of indeed anything is finite, limited by metabolic and neuronal factors and must somehow find an output which does not require infinite computations, which is a necessary consequence well accepted in predictive coding and both the infomax and free energy principles (Friston et al., 2012). Essentially, the brain of the subject is here forced to prefer an error of one kind over an error of another kind. I would suggest that forward and backward connections mediate this by choosing a predictive model which silences the most important error and accepting any remaining errors. What kind of error is acceptable will depend on the context, in this case how a question about semantic content is construed. For Ferreira, the actual strict semantic content of a sentence, even if implausible, was “correct.” Let us consider the example sentence ‘No head wound is too trivial to ignore.’ Ferreira’s interpretation would look as in Figure 7.

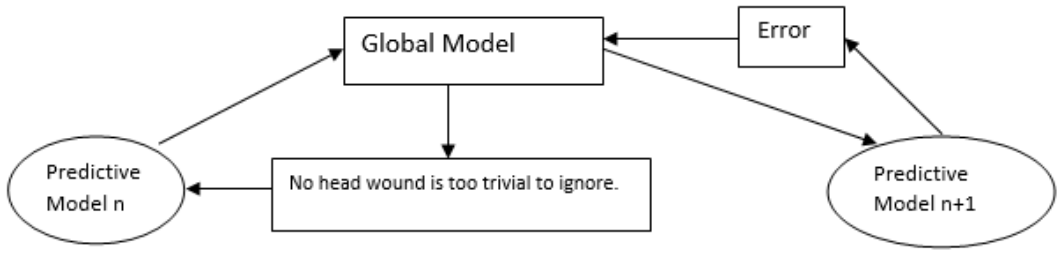

Figure 7. A strict semantic interpretation of the input phrase leading to an error between the new predictive model n+1 and the global model

In this case a participant would strictly interpret the sentence, using their knowledge of the word meanings from their global model, and form predictive model n+1. Predictive model n+1 however now clashes with the participant's global model, which likely does not contain the idea that any head wounds should be ignored, and certainly not all head wounds.

For that participant, it is quite likely that Fillenbaum's suggestion is more appropriate and that correctness is measured against which statement might more accurately reflect the supposed situation being communicated, in which case the normalized version is "correct." The participant will see the error as stemming in the actual communication, which is the sentence being uttered. This would look as in Figure 8.

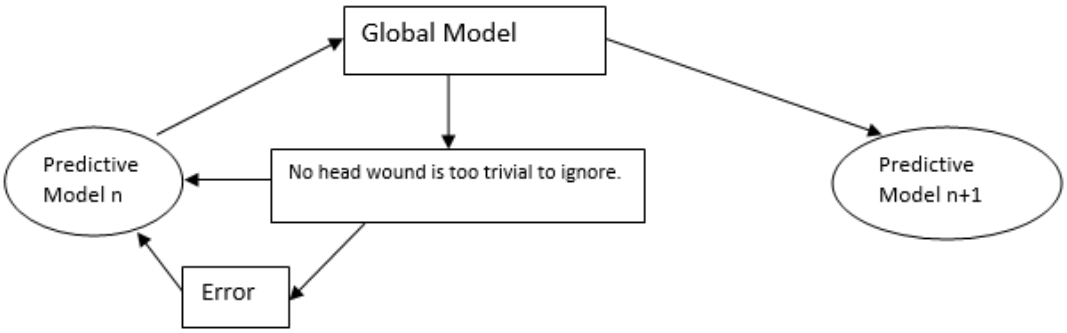

Figure 8. An error is elicited between the perceived utterance, and its semantic meaning when compared to global model expectation, leading to new predictive model n+1

In this case, the participant infers that the speaker meant to say 'No head wound is trivial enough to ignore' or something similar. An error is formed within the utterance, and new predictive model n+1 contains the interpretation based on the normalised sentence.

We can see this in more detail considering another dual processing approach proposed by van Herten et al. Their theory rests on the principle that cognitive self-monitoring of errors during speech production also forms an intrinsic part of speech perception (van Herten,

Chwilla, & Kolk, 2006). Therefore, different ERP effects, which produce a measurable increase or decrease in voltage within the brain, measureable by electrodes on the scalp, have been observed during language comprehension, which the researchers take to be signs of a conflict between two processes of comprehension. They also distinguish between a classical parsing algorithm and two heuristics: the noun-verb-noun, or NVN strategy also mentioned by Ferreira and a plausibility heuristic, which attempts to combine the lexical items of a sentence in the most plausible way (van Herten et al., 2006). The curious difference between ERP responses which must be incorporated into the notion of monitoring and repair is that semantic anomalies tend to elicit a so called N400 effect, while others do not. The N400 is a strong negative brain wave effect presenting around 400 milliseconds after being presented with a stimulus, using non-invasive electrodes attached to the scalp. The N400 effect is well documented in a number of studies of unusual or unexpected semantic anomalies (Nieuwland & Van Berkum, 2006; Sanford & Emmott, 2012; van Herten et al., 2006). When comparing plausible sentences to implausible sentences however, no N400 is measured, but instead a P600, a large positive brain wave effect measured around 600 milliseconds after the stimulus sentences are received (van Herten et al., 2006). This is taken as evidence for the fact that there must be a process of conflict monitoring at work, which does more than simply recognize anomalies.

The assumption made is that algorithmic and heuristic processing happen in parallel, leading to three different outcomes: first, both algorithm and heuristic deliver a plausible outcome, leading to no effect of either kind. Second, both routes deliver an implausible outcome, leading to an N400. Third, the algorithm delivers an implausible outcome, but the heuristic can deliver a plausible one, leading to a conflict and repair which elicits a P600 effect (van Herten et al., 2006). These were tested in several experiments using sentences with unacceptable agent-patient pairings, and garden-path sentences with multiple anomalous conjunctions. The results generally confirmed this prediction, but also led to a surprising fourth situation: some sentences which were supposed to contain non-reversible anomalies, such as “The apple that climbed the tree looked juicy” elicited both the expected N400 effect and an unexpected P600 effect. The hypothesis was made that the P600 effect occurred because although the algorithm route is unable to come to a plausible outcome, the heuristic route was able to make plausible sense of portions of the sentence, like climbing a tree (van Herten et al., 2006, pp. 1188–1189). This was tested in a second experiment using sentences carefully constructed to have anomalous subject-verb relations but plausible verb-object pairs. This experiment confirmed the hypothesis and showed that

anomalous sentences containing plausible subsections yielded P600 effects and reduced or even no N400 effects (van Herten et al., 2006). The conclusion is that the plausibility heuristic can overrule both the algorithmic and NVN heuristic strategies, giving the preferred late response to the most plausible interpretation even at the cost of actual lexical meaning of a sentence (van Herten et al., 2006, p. 1194).

Friston himself gives a very limited example of this effect, considering a similar anomaly when reading the sentences “Jack and Jill event up the hill. The last event was cancelled.” Most readers normalize the first sentence to read “went” instead of “event.” The very short explanation accompanying this is that a strong semantic prior favours the interpretation of the more plausible “went” and accepts an error at the visual level of the word recognition in favour of minimizing a larger error at the semantic level (Friston, 2002, p. 237), see also (Friston, 2003). I think this is correct, and we can apply the same analyses as I have offered of the phrase “No head wound is too trivial to ignore” above. All of these examples have occurred at the semantic level, at the interpreted meanings of utterances and words, irrespective of our ability to perceive and interpret them at a symbol level, and this highlights the second hierarchical level of the processing as I have introduced it in Section 2.1, repeated here for clarity as Figure 9.

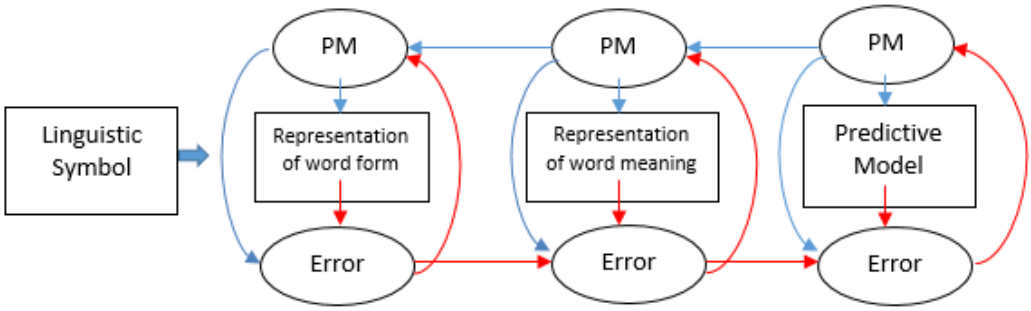

Figure 9. The prediction and error cycle leading from perception of symbols to predictive models

The conflicts discussed have been at the representation of word meaning level, and have not required any argumentation about the validity or interpretation of the word forms.

In this Section, I have discussed the notion of word meanings from the perspective of cognitive grammar and linguistic research. I have argued that the idea of construals of cognitive grammar is fruitful and can be well integrated into the predictive model theory framework. Words with multiple meanings are resolved by predictive models which select based on context. I have also introduced the effect of pragmatic normalisation, and

suggested that a predictive model analysis is necessary to explain why participants respond with normalisations in context, and why I believe it is wrong to speak of shallow processing or of a speaker making a mistake when normalising an utterance. This leads on from the notion that rather than a simple one-way system of words and references, we activate a complex web of knowledge surrounding word forms and word meanings, then use predictive models to select against an error in context to arrive at an interpretation. In the next section I will discuss conflict monitoring and mismatch effects, and show how predictive models can explain these effects.

2.3 Meaning and Mismatch Negativity

In this Section I will introduce the idea of conflict monitoring. I will then argue for a predictive model analysis of this phenomenon, to explain why we can observe what appears to be a top-down process in which a speakers monitors their own utterances for errors. My argument will be that this is part of regular predictive model processing. I will also discuss mismatch negativity (MMN) in more detail and discuss the role of my framework so far in explaining these effects.

Let us turn to conflict monitoring. Most of the conflict monitoring theory above fits very well into the framework, and can deal with a few difficulties a monitoring strategy alone cannot. A mild mystery of the monitoring approach is why there are measurable error effects when someone is told they have made an error, even if they have not (van Herten et al., 2006). This account actually makes sense if we consider the monitoring process to be high level predictive models which are continually running. In this account, errors would be elicited by perceptual self-monitoring on a lower, speech-production level but also at a higher conceptual level by being informed of an error, causing internal self-monitoring. Similarly it is not clear from the monitoring account when or why the monitoring process should be using an algorithmic and a heuristic process, especially given metabolic constraints. Instead I would suggest that it is always running, as part of overall predictive model processing.

The perception of strings of linguistic symbols leads to a process of recognition and prediction in the visual areas of the brain (or auditory areas for speech of course). Word forms are stored as discussed in section 2.1, resembling perceptual symbols, or cell

assemblies specialized in responding to specific perceptual patterns (Barsalou, 1999; Pulvermüller, 2008; Wennekers et al., 2006). These perceptual symbols of course have not only recognition models attached to isolated symbols but also to association of certain symbols which co-occur habitually. Because of these structural associations, it is likely that some errors of semantic anomaly actually already occur during perceptual processing, if words which do not co-occur statistically may have developed inhibitory connections (see Section 1.2). Predictive models become active to attempt to predict upcoming word forms and to match symbols to learned forms and fill in any gaps if there is poor visibility, distraction or any other interference to the perceptual input.

The initial predictive model attempts to match the incoming sensory data to known things. If the sensory signal is good, easily recognized linguistic symbols are immediately dealt with by a predictive model within forward connections alone. Those same predictive models however are likely to contain connections to other symbols because linguistic symbols tend to occur in groups. A predictive model will ensue which contains a prediction of the next symbol, which if successful will greatly ease processing of the rest of incoming sensory data. The prediction may draw upon certain parts of the brain which deal with sensorimotor reactions if the prediction simulates producing the sound to form a prediction of what a speaker is saying, or how to write or otherwise shape the word form (Barsalou, 1999; Garrod et al., 2014; Pulvermüller, 2008; Pulvermüller & Preissl, 1991). If the prediction does not match the word form identified from the newly incoming signal, there is a clash. This is where mismatch negativities (MMN) come in. If the clash is minor, because there is interference, or a word is misspelled, then a new cycle of errors and prediction through lateral connections on the same level will be sufficient to repair it. If there is a large error however, because the word form is entirely unexpected and not previously associated with the preceding form, we get a very large error. Errors due to rare stimuli are associated to mismatch negativity effects in auditory and other areas within Predictive Coding, and there is good reason to believe that they represent an inability to cancel out prediction error (Friston, 2005).

The effect typically occurs when there is an immediate semantic anomaly, such as a wrong or entirely unexpected word. The anomaly must also be of a sufficient severity that it cannot be cancelled out easily by a new predictive model, and we can see this in the distinctions of sentences which elicit the effect. Nieuwland and Van Berkum demonstrated the effect using sentences with implausible or impossible agents, such as a yacht consulting a psychotherapist (Nieuwland & Van Berkum, 2006). Van Berkum et al. also showed similar

N400 effects when creating sentences in Dutch which caused listeners to predict nouns of a certain gender, then presenting them with target words which clashed with the predictions (Van Berkum et al., 2005). For van Herten et al. it was the kind of sentence they termed nonreversal unacceptable, i.e.: "The apple that climbed the tree looked juicy," which elicited notable N400 effects, followed by a surprising P600 effect of positive electrical activity presenting to the electrodes. I believe this is because the sentence is the kind which cannot be repaired by a normalisation or by taking a strict semantic interpretation. Interpretation would run as in Figure 10 below. The sentence brings up an error as apples cannot climb trees, violating the one-to-one principle by giving an action that this agent cannot perform. There is however no information within a hearer's global model which can resolve this, and no real alternative meaning that can be taken into account to resolve this via normalisation, so a new predictive model still containing the error is formed, which now continues to conflict with the global model.

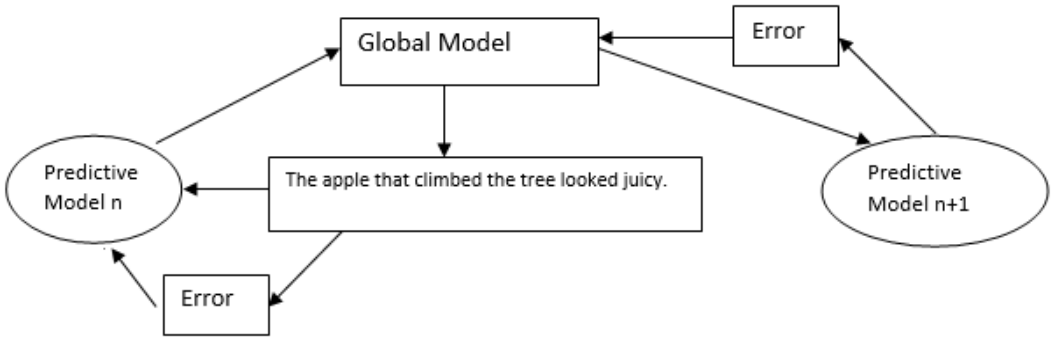

Figure 10. A semantic error cannot be resolved, and leads to a new predictive model forming, which propagates the error signal to the next level of the hierarchy.

The P600 is likely caused when further error suppression is attempted, and a hearer is able to match fragments of the utterance into the global model. We can integrate apples being juicy, apples being on or near trees, and climbing trees, but these actions cannot be resolved in the same predictive model without adding another agent. All of these are examples of prediction errors which were purposely introduced by the researchers, causing an error signal which their research subjects could not solve using a prediction, nor repair using context. Van Berkum et al. believe this to mean that predictions about upcoming words are routine and effortless, which is precisely what I am suggesting as well (Van Berkum et al., 2005, p. 451). What makes the account interesting, and what is also crucial to my point, is that the effect of this very same error could be reduced or even disappear entirely by placing it in an appropriate context.

When the abnormal condition was reversible, as in: “The ladder that climbed the painter suddenly fell,” in which there is of course an anomaly, there were nevertheless no significant N400 effects present (van Herten et al., 2006, p. 1188). This was reasoned to be because the anomaly can be repaired by reversing the agent and patient, with which I fully agree. This conflict can be solved with a normalisation as in Figure 11, in which the error in the sentence can be normalised by a new predictive model containing the interpretation that the speaker of the utterance in fact meant ‘The painter that climbed the ladder suddenly fell.’ Now a further repair is not necessary, and neither the N400 of the error propagating up to global model integration, nor a P600 for any partial error suppression can be caused by the utterance.

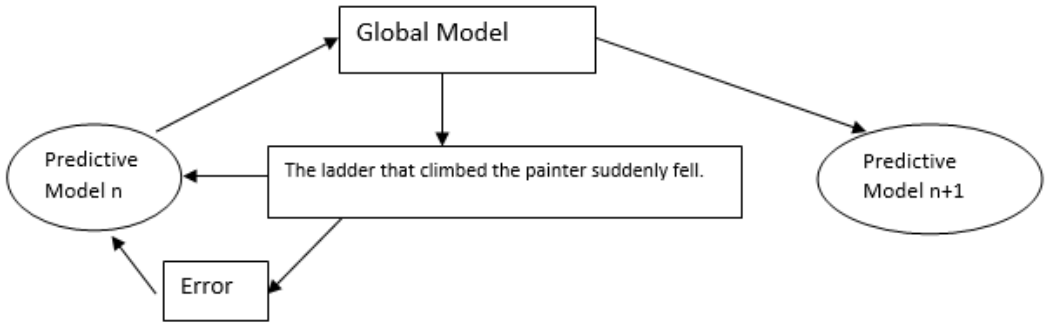

Figure 11. An error in the perceived utterance is repaired by a normalisation in which the word order is changed, forming new stable predictive model n+1

Similarly, when placed in appropriate context of a story situation, Nieuwland and Van Berkum found that the N400 effects to unexpected words were greatly reduced and eventually disappeared (Nieuwland & Van Berkum, 2006). The same reduction of N400 effects to unexpected words was shown using context of everyday events. If a word was locally anomalous within a sentence but related to the kind of event the sentence described, N400 effects were greatly reduced (Metusalem et al., 2012). We can draw several conclusions from this circumstance.

The N400 effect cannot be caused by a simple clash of word forms, where word form co-variation is learned statistically. If that were the case, then there should be no possibility of reducing an N400 effect by variation of context, nor of inducing one by manipulating the context. The definition of what an error is always stems from the definition of what the expectation is, therefore a purely statistical expectation of learned word forms would not be able to be contextually sensitive if there were not a higher hierarchical step giving the

context. Likewise, if there were not multiple steps of prediction, errors and renewed error suppression, it should not be possible for utterances to elicit multiple different effects.

While we cannot say for certain where precisely in the process they occur, these ERP effects are further good evidence for viewing the interpretation process in terms of the predictive model theory, as in Figure 12, repeated here from the previous Sections.

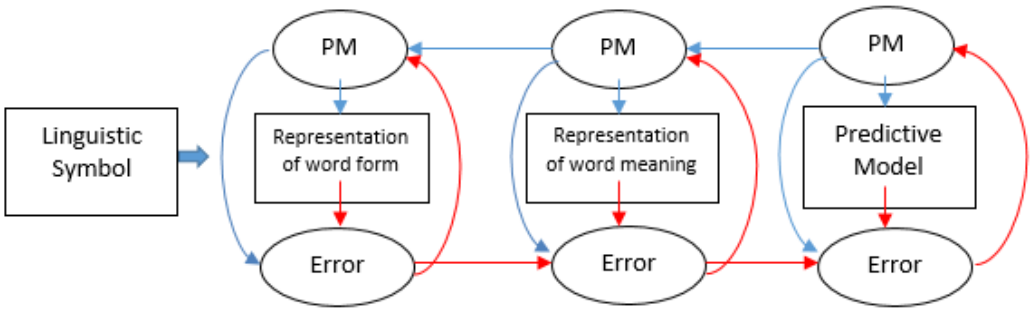

Figure 12. The prediction and error cycle leading from perception of symbols to predictive models

A clash of expected word meanings during sentence anticipation and prediction can cause the effect, while a possibility of assigning a meaning which overrides the strict interpretation of word forms but which can integrate to the global model successfully is capable of repairing it. If both processes fail, then other, hierarchically higher processes attempt to suppress the remaining errors. This is another reason why I consider them heuristically separable steps of processing. Of course recognizing word forms initiates the process, and error signals caused by word forms are sent up the hierarchy, but clearly a new, more complex system which is independent of the arbitrary nature of the symbolic forms sits on the higher levels of this hierarchy. Most importantly, it makes the most sense to consider these as a hierarchy, because in all of the above evidence of both pragmatic normalization and MMN effects there is a clear conclusion emerging. *Derived meaning trumps perceived linguistic symbols and the global model trumps derived meaning*. In a conflict where the linguistic symbols cause an error, either by an anomalous order or an unexpected symbol in the chain, the process assigning meaning to the symbols will assert any plausible interpretation it can and suppress or accept the "lower" errors. If neither system is able to overcome the error, then a noticeable N400 effect shows the persistent error propagating up the hierarchy.

It also makes sense that repair effects such as the P600 should be measured later than the N400, as the repair occurs higher in the hierarchy and thus chronologically later, and makes

it far more plausible why there should indeed be a phenomenon of partial repair as reported by van Herten et al.: some errors within sentences may be too large to be overcome, but the predictive model sent backward will always attempt to suppress it anyway, and quite naturally deal with any subsets of information that can be correctly predicted. In reversal conditions, such as the ladder climbing the man, the predictive model can reverse it entirely and suppress the entire error. In the example of an apple climbing a tree, the situation cannot be reversed of course but as the researchers suggest, it is likely that climbing a tree forms a sense that can be used to partially suppress the error (van Herten et al., 2006). We can speculate that the studies of Nieuwland and Van Berkum and Metusalem et al. may well have found P600 effects when introducing appropriate contexts to other target sentences. Rather than finding the partial fits within the sentence, here the information gained from the contextual statements introduced elements that allowed predictions to lessen and finally suppress the error, which should also have caused some form of repairing effect. Finally, the hierarchical account can mediate the positions of van Herten (2006) and Nieuwland and Van Berkum (2006). The first group believe there to be a parallel monitoring process (van Herten, 2006), which repairs itself in two distinct steps, while the latter prefer a single step model (Nieuwland & Van Berkum, 2006). I offer a compromise: it is a single process, but one entailing multiple hierarchical steps. This offers the best possible explanation for the observed outcomes, and can easily deal with how the N400 effect is influenced, while remaining plausible for the effects of self-monitoring of speech production as well as perception. There is also no need to distinguish between processes for local or wider situational context, as the predictive models employed by the hierarchy naturally incorporate any contextual information which can suppress errors.

In this Section, I have discussed the effects of conflict monitoring and mismatch negativity. I have argued that the processes underlying both can be explained well using the predictive model theory framework. I have also suggested that certain well documented mismatch negativities may be clues to certain hierarchical stages in the predictive model processing. Now let us consider what the hierarchical levels to do with meaning might contain and where their limits lie, as well as how we may define objects and concepts both within predictive models and the global model.

2.4 Meanings and Concepts

Having made a formal divide between word forms as symbols, and meanings as associated to these, I will in this section discuss more closely what "meaning" is from a theoretical standpoint and how the meanings we hold in our minds relate to the world in more detail. The sections leading up to this have set a foundation, but I shall delve deeper into cognitive grammar, perceptual symbol theory and my own framework in order to show how meanings are processed, associated to symbols, and what this means for our understanding of the world, and things beyond it.

A very good way to start describing how a conceptual meaning might work in the brain is to return to Barsalou (1999). While the criticism that he did not sufficiently differentiate words from conceptual meanings stands, his description of the conceptual meanings as such is quite convincing. One of his earliest and purest formulations of how perceptual symbols arise is the following:

> Once a perceptual state arises, a subset of it is extracted via selective attention and stored permanently in long-term memory. On later retrievals, this perceptual memory can function symbolically, standing for referents in the world, and entering into symbol manipulation. As collections of perceptual symbols develop, they constitute the representations that underlie cognition. Perceptual symbols are modal and analogical. They are modal because they are represented in the same systems as the perceptual states that produced them. The neural systems that represent color in perception, for example, also represent the colors of objects in perceptual symbols, at least to a significant extent. (Barsalou, 1999, pp. 577–578)

Prima facie, this is a perfectly usable summary of how "meaning" as causal representations come about. We learn how to speak, read and write by having perceptual experiences which are stored and reapplied as recognition models in later experience. This process does not just cover aspects of learning linguistic symbols, but also general information about the world. We can follow Barsalou as above on this. Perceptual experiences are stored and used for future recognition models, with specific information or at least specific preferred responses becoming learned through Hebbian learning. As the incoming sensory signal is dealt with initially by a cortical area which is specialized for this signal (necessitated further by the fact the nerve endings coming from sensory organs connect at the brain at a specific place), it is logical to think that initial predictive models will arise within the same area. Any direct Hebbian learning which has to do with recognizing the same signal again will also happen within this area, leading any recognition models which may be used symbolically to also be stored here, but likely also in other places, which we will discuss shortly.

This view requires two caveats. One is addressed by Barsalou in the form of subsets. It is always necessary to see that any stored cognitive information about the perceptual experiences we have of the world is a small subset of the actual information it is possible to gain. According to Barsalou, this subset is chosen through selective attention, which focusses on the coherent aspects of what is perceived, reinforcing the representative neurons in engaging in Hebbian learning (Barsalou, 1999). This viewpoint is based on the exact same principle of neural representation as I understand it, and also aligns well with predictive coding, as Barsalou assumes specific neural populations which represent sensory information, whose activation patterns are stored for future recognition (Barsalou, 1999). Selective attention is a difficult process to deal with, but it has been associated to better long-term memory storage. In Predictive Coding it is associated with increased precision in dealing with prediction error, by modulating the strength of error signals and the resulting response to them (Clark, 2013; Friston, 2002). Given this, it is quite plausible to think that recognition of objects and future recognition rests on such a process. The most salient, most attended to aspects of a sensory signal are those which cancel out errors, and integrate well into the global modal which can lead to them being stored for later use as recognition models for similar items.

The second caveat is that the process of learning symbols proposed by Barsalou (1999) presupposes that selective attention focusses on salient aspects of objects and events, with which I agree, but does not explain how perception knows what objects and events are or what makes them salient. Definitions of “object” and “event” are notoriously difficult. In the literature, they are rarely defined. None of the sources cited so far in this work have given a definition of object, or event. There are some definitions of events in literary studies, the most notable to the present endeavour coming from Ryan: “Events are perfective processes leading to a change in truth value of at least one stative proposition,” (Ryan, 1991, p. 124). Sadly, for the attempt of assigning a perceptual recognition model of an event, this is rather useless. Even the Oxford Dictionary is not particularly helpful here, defining objects as: “A material thing that can be seen and touched,” (“object - definition of object in English from the Oxford dictionary,” n.d.). Event meanwhile is defined as: “A thing that happens or takes place, especially one of importance,” (“event - definition of event in English from the Oxford dictionary,” n.d.). Objects then are things that can be seen, and events are things that happen or take place, yet presumably both can actually be seen, or be of importance. Someone wishing to break the circularity of these definitions by resorting to the meaning of “thing” will be disappointed, as it is defined as: “An object that one need

not, cannot, or does not wish to give a specific name to," ("thing - definition of thing in English from the Oxford dictionary," n.d.). The result is not very satisfying. *Objects are things, events are things, and things are objects*. This circularity is rife in the literature, in which no real definition breaking such circles is ever found. It is also found in German, where both objects and events are defined in terms of an "Etwas" (something) whose definition is "A not clearly defined being or thing" ("Etwas - definition of etwas in German from the Duden," n.d.). It would be nice to be able to say more on the subject, but the only real conclusion which can be gleaned from the evidence is that it appears that our concepts of "object" and "event" are, among others, epigenetic priors which the brain uses to classify certain inputs into beginning-middle-end structures, or singular object categories as necessary during real-time processing.

This view is consistent with a fundamental knowledge structure called an "image schema." Image schemas can be summarized as knowledge structures based on the basic physical structure of both the human organism and the world, which are basic, preconceptual and meaningful at an unconscious level. Among them are listed such categories as containers, motions and lines of motion, force, contact and more (Hampe, Grady, & ebrary, 2005, pp. 1–2). I believe that "Object" and "Event" need to also be considered in these lines. These conceptual priors are what we associate to the word forms defined above, which allows a human being to overcome the circularity and instead function while automatically applying the prior definitions to the world. This fundamentally basic nature of such concepts may also play part in our difficulty with defining them: there are really no more basic concepts to reach for in attempting to do so. It is possible that these priors in a sense represent not any cognitive limitation, but a biological one. The brain is made up of neurons which fire and become active, whereby a certain set amount of neurons firing constitutes a representation (see Section 1.2). Such a set amount of firing neurons has the very same properties: a finite amount of bounded neurons is firing, representing a finite and bounded object or set of objects. Given this neuronal structure, it is not necessarily possible for the human brain to conceive of objects in any different way, as flowing neuron activations would not give clear boundaries for representation. Similar, the spatio-temporal nature of neuron firings, with a clear temporal and spatial beginning and an end, is likely what gives shape to our prototypical event structure, and would be equally difficult if not impossible to overcome. It is how our brain is built, and it has proved adaptable and successful enough to evolve to its current state. From the standpoint of individual success in understanding and manipulating one's direct surroundings, it seems quite logical that this is a highly useful

brain structure, as our definitions of objects and events allow direct causal reasoning, while dealing with limited enough information about the world so as to not be constantly overwhelmed.

With that said, having an epigenetic prior in place for what a singular, worldly object is, and that series of happenings or changes in things can be construed as event chains, perceptual symbol theory becomes much more useful. The perceptual areas of the brain attempt to naturally look for objects and changes within objects, and selective attention which focusses on edges or other salient qualities leads to the learning of recognition models. The stored and learned models are not static structures that always activate the same way, but they are dynamic and may be only partially activated depending on context (Barsalou, 1999). This is also quite plausible and fits with the idea of contextual modulation. If a predictive model cannot perfectly account for the incoming signal, then partial fits may still be used, with resulting errors propagating forward. The nature of the models stored may be generic, or determinate, and Barsalou makes a good argument for the fact that perceptual symbols are quite capable of being abstract because the neural representations do not need to factor in exact or specific aspects such as length, size, or number; therefore when conceiving of a tiger it is not important how many stripes it has exactly (Barsalou, 1999, pp. 584–585). This abstractive ability is aided by the fact that concepts can also be combined in novel ways to consider objects that have not been perceived (Barsalou, 2003). Information about the objects is not just stored in one modality, but of course among multiple sensory modalities, of which Barsalou lists seven: the primary senses, vision, audition, haptics, olfaction, gustation and in addition proprioception and introspection (Barsalou, 1999, p. 585). Taking only this view, and Barsalou's more basic formulation above, that the perceptual memory can function symbolically, I would like to propose a slightly more radical, alternative proposal for how conceptual meanings are stored and related to language.

Let us assume that as I have mentioned, "object" and "event," as well as "attribute" and "adjective" relations are epigenetic priors. They form basic categories that the brain can choose to apply to incoming sensory signals, at any level of the hierarchy, but usually straight away. Following from the tenet that some existing knowledge is needed for the recognition of objects in perception (Helmholtz, 1962), then the very basic knowledge needed, before knowing a single object, is that there are objects and that our job in understanding incoming sensory signals is to identify what the objects in the world are. This is an elaboration of Predictive Coding and the free energy principle. If the basic goal of the

brain is to correlate causes to sensory signals, then the most basic formulation of a solution would be: "I perceive an object because there is an object." *The basic business of object recognition in the brain is thus the supposition that the cause of any incoming signal is the world, and the cause of our perception of separable objects is the material existence of such objects in it.* In direct terms, when we perceive something in our environment, the standard predictive model is that we are perceiving objects and that the sensory signal and any errors with it can be cancelled out by identifying what the objects are and what their attributes are.

The most salient of such attributes are stored as Barsalou (1999) suggests, with Hebbian learning forming specific recognition models. Because attributes overlap, our store of possible attributes very quickly becomes much larger than our store of actual objects, and also covers far more modalities. The nature of something being an abstract attribute or a defined object becomes obscured because we define objects by what attributes they have, without ever actually defining the objects themselves in any uniquely identifiable way. In this sense, Barsalou's (1999) argument that object knowledge does not have to be specific follows quite naturally, because the object is defined by specifying a catalogue of attributes which we know or believe it to have. Thus we can conceive of a tiger with an indefinite number of stripes, or perhaps twenty stripes, or forty, or even just one, because we have already attributed objecthood to the tiger. Stripes themselves, as an abstract concept come from the objectification of one of them, stripe, which can then be elaborated as needed when applied to another object. This dual object-attribute nature permeates many of our concepts, even very abstract ones, and contributes to the enormous flexibility of our conceptual system; consider things such as red-redness, nothing-absence, time-duration. *Therefore, the storing of objects and their attributes, their perceptual symbol knowledge, is not separate and not a case of learning both abstractions and concrete examples, but rather a field of knowledge on which we super-impose generality or specificity by applying the basic priors of objects to certain collections of them. That is, our definition of objects is given by specific collections of attributes we associate with them. The consequence is that interpreting objects as objects is a dynamic thing, and must be negotiated online when deciding on the interpretation of an incoming signal.*

Representing something as an object can work as described by Barsalou's "simulators." Simulators are multi-modal systems in the brain which can represent concepts (Barsalou, 1999, p. 586; 2003, p. 521; 2009, p. 1282). As perceptual knowledge about something is gained, for example a car, selective attention focusses on aspects of experienced cars and

begins to associate them to form a frame of what a car is, along with a multitude of different attributes cars may have, such as colours, frame shapes, tire shapes and so on (Barsalou, 1999, p. 590). By activating and combining these different attributes, the frame can be used by a simulator to represent a car of any kind, creatively combining stored features with each other and changing their size and orientation. These features are also context sensitive, as certain constraints which are either learned or introduced by a specific experience. If no direct context is available, or a constraint is fairly weak, then a default may dominate the process (Barsalou, 1999). The resulting simulator is a distributed brain system that represents the concept or category (Barsalou, 2009). This simulator can equally well be understood as a predictive model system, which combines and modulates stored attributes as needed into the representation of an object. I would add to Barsalou's account that the represented objected can be our prediction of what we are perceiving, what we are imagining, or a response to a linguistic symbol, regardless of whether any supposed object actually exists, or whether we are right or wrong about supposing so.

The last point is important, because it is the cause of error in many illusions and confusions about the world, which are caused by our predisposition to identify bounded collections of impressions as objects, and assuming them to be the cause of our perception. Consider once more the example of the morning and evening star from Section 2.1. Many observers thought for many years that there were quite clearly two unique objects observed in the sky. It is hard to blame them, as the perceptual impression given by these "objects" was rather unique: the collections of attributes associated to them are different, and the natural assumption of our object prior is that different attributes equal different object. The conclusion stands. Perceptual symbols do not necessarily refer to real objects. So what are they symbolic of? They are symbolic of our representations of what the outside world is like. The brain simply has no access to physical reality, only to indirect sense impressions and the basic assumption that there is a world and objects within it. When faced with sensory information about specific attributes which it can constrain into an object, it will do so, even if that object was already perceived in a different way, or if there is no object. The entailment of this conclusion is that since language as a symbolic subsystem refers to our conceptual knowledge, and thus to perceptual symbols, language also does not refer to the world directly. It refers to our representation of the world and the represented objects we believe to exist in it, or when reading a text, to the objects we believe to exist as described within it. We then apply these beliefs to the world, or the text, in the hope that they will match future inputs. In sections 2.1 and 2.2 I began this argument by differentiating

between words referring to objects directly or to concepts or the like. This now is the culmination of the way this type of reference must work. In essence, direct reference from a word to an object, that is from our stored knowledge and linguistic symbols to the object, is a matter of matching concepts and input internally, then relying on these to match within the context they are encountered in, and in a communicative situation also relying on them matching in the mind of the listener or reader. In a fictional setting when we are reading a text, this difficulty is mitigated in the sense that we do not necessarily receive further input to verify if our predictions work or not, as a text may leave this open or never address it again. In these cases a reader is free to assume any matching that satisfied their global modal expectations. If the fictional object or concept in questions does appear again in the text, then the same process as when matching predictions for a worldly object or concept will apply. I will discuss specific expectations regarding literature far more closely in the following chapters.

To return to word meanings, the field of possible meanings which can be attributed is exactly the body of conceptual and perceptual knowledge we learn through our now slightly modified interpretation of Barsalou (1999, 2003, 2009). Through salience and attentions, Hebbian learning allows us to associate attributes into groups, and predictive models generalise them into objects where needed. Specific objects become learned as they are experienced and given names. Importantly, following Barsalou (1999, 2003, 2009), we are not just talking about outside sensory inputs, but all things experienced through proprioception and introspection. Concurrent language use as we learn these things also establishes associations between words and our combined conceptual/perceptual knowledge about the world. Words become concepts which are associated to parts of this knowledge base in order to represent it symbolically for communication and for more structured internal processing. Learning word forms themselves, as phonological and graphological symbols, follows the exact same rules, and it is nicely shown that there is a similar hierarchical nature behind both of them in functional grammar (Halliday & Matthiessen, 2013). We will assume for now that we are speaking about a reader who already has a fully developed language and the amount of experience and general knowledge any adult might.

World knowledge and linguistic knowledge interface this way to create the joint structures that represent word concepts, by transforming knowledge of the world into linguistic meanings (Halliday & Matthiessen, 2013). This is just the interfacing stage, and it is quite possible for world knowledge to exceed our linguistic capabilities and for linguistic

representations to exceed our world knowledge. The first happens when we discover things we cannot describe yet, the second when we learn something new encoded through language, i.e. reading a textbook. The way functional grammar defines this effect however relates very well to my own wording, classifying the essential word classes as the names of knowledge chunks: nouns are the names of entities, verbs are the names of processes and adjectives the names of qualities (Halliday & Matthiessen, 2013). Such a classification follows naturally from our automatic tendency to perceive attributes, group them into objects and perceive changes between them as processes. However, just as conceptual knowledge in this form has variation and can be simulated or construed in different ways, words do not necessarily have one set meaning but can also vary their meaning in usage. Basic word functions such as “subject” can be seen as operating on at least three different levels: psychological Subject, grammatical Subject and logical Subject, which correspond to not simply three different aspects, but rather distinct meanings which can be represented by one word or split apart within a sentence (Halliday & Matthiessen, 2013).

The different subject types can correspond to different meaning functions of clauses. In a clause as message, the subject is intended to impart information. In a clause as exchange it functions as the warranty of an exchange. In a clause as representation it is the actor of a process of human experience. These three functions run throughout the English language and co-exist without being antagonistic; a sentence can contain all three meanings or one of them (Halliday & Matthiessen, 2013). This echoes the structure of world knowledge built up as perceptual symbols. Subjects, both as physical objects and as social entities, can be perceived in different ways. A beholder might see an agent, refer to an individual and so on, and we can also postulate that in language these levels overlap and we are able to conceptualize being informed of someone doing an action, as well as this being intended as a truthful proposition.

On the question of how these words and word forms interact with this knowledge, I follow Croft and Cruse (2004) as well as Langacker (1986) who describe the relationship as one of access nodes into a network. They assume, correctly I believe, that rather than only parts of our knowledge, words activate all knowledge that might be associated to them. While the nature of world knowledge is never addressed in cognitive grammar literature, I believe that this network access metaphor fits rather well onto the view of basic world knowledge I have described thus far. Experiencing a sensory signal recognized as a word form leads to a prediction that a specific word is intended to be symbolized, and this in turn leads to higher level predictions of which parts of our knowledge can be used to interpret it. This leads to

the question of conceptual boundaries and categorization. Of course we do not simply have a mess of knowledge clustered around the brain without a system; it would be difficult to fulfil any object/attribute relations meaningfully without a categorization system. Defaults, defined by the strongest connections formed through Hebbian learning can account for this to a degree. The boundaries of any categorization however must be flexible to account for new information or novel combinations of conceptual meanings. Croft and Cruse (2004) resort to what they call "dynamic construals" for this purpose. They point out that in various responses, categories can shift with context to include or exclude category members quite fluidly. Students asked "Is a cyberpet a real pet?" overwhelmingly respond "No." If they are told that a psychologist advises parents of a problem child with: "I advise you to get her some kind of pet – even an electronic one might be beneficial," then the same students find no anomaly with the utterance, despite the fact that the categories for "pet" overlap in both statements (Croft & Cruse, 2004, p. 94).

While classical theories of conceptualization which rely on feature lists or specific linguistic markers fail at being able to explain such shifts of category meaning, the principle of dynamic construal says that there are no such structures. Instead, meaning and conceptual boundaries are decided "on-line" in actual use, utilizing clues from the actual linguistic symbols but also context and other activated knowledge (Croft & Cruse, 2004). The important caveat is "that concepts are not necessarily equatable with contextually construed meanings, or, as we shall call them, **interpretations**," (Croft & Cruse, 2004, p. 98). They leave the neural underpinnings to this somewhat unclear however:

> An interpretation resembles a picture in that it is not susceptible of finite characterization in terms of semantic features, or whatever. Any features are themselves construals. Of course, a meaning must in some sense have a finite neural representation, but the elements out of which the representation is composed are more like the pixels underlying a picture on a computer screen: the resulting experienced picture is a Gestalt and so is an interpretation. The nature of this experience is still mysterious. (Croft & Cruse, 2004, p. 100)

While this statement is not further backed up with any causal argument or evidence regarding what this interpretation then actually is, we can explain it plausibly with predictive model theory. Using elements from the knowledge base which are associated to the index marker that is a word, an interpretation is the contextually salient meaning assigned to it by backward connections. *An interpretation is contained in a predictive model constructed for that word or the utterance containing it in that context, and as such it is always going to be unique to that exact context*. This can explain how we can have relatively stable meaning structures thanks to Hebbian learning yet flexibly assign meaning to odd or novel statements. The way in which smaller structures of attributes and

perceptual symbol knowledge are combined within the predictive model resembles the metaphor of pixels forming an image rather well, and mirrors objectification. A predictive model construes a word meaning through selecting the right combination of stored attribute knowledge, just as it construes an object in perception or a concept in introspection. Sometimes these overlap, such as when a concept or category must be construed to attempt to fit a word interpretation.

More precisely, the driving signal coming in through forward connections demands a response from the knowledge structures associated to recognition models for it. In this case the signal is interpreted as a word form, which is in turn interpreted as a symbol, associated to neural populations encoding perceptual symbol knowledge. The predictive model must now find the best way to "explain" and cancel out this signal, and begins modulating synaptic connections. It does so by dynamically altering the plasticity of the connections, inhibiting some neurons activated by the driving signal which do not fit the predictive model while facilitating those that do, and this is the biological basis of the dynamic category construal Croft and Cruse describe. Because these boundaries are not inherent to the knowledge base but instead mediated by backward connections using temporarily mediated plasticity, they can be re-established differently when a word is encountered in a different context, or even on-line within the same interpretation if an error with the predictive model is noticed, or perhaps if several models fit.

In summary, the way that we can define and present the concepts of "events" and "objects" within predictive models is as follows. When we receive streams of input through our sensory organs, these are bounded and interpreted within predictive models which contain priors that naturally categorize the information to have beginning, middle, end, and bounded objects within them. An event is perceived as following the one-to-one principle (Section 1.4) wherein one agent performs one action at one point in time and at one location. The smaller unit of this structure is the object – objects are collections of attributes which share a spacial and temporal location. Thus we always attempt to bound objects into specific finite shapes and areas which share a particular set of attributes. But this is part of our predictive model, and not necessarily a truly accurate definition of the reality. In order for our perception and recognition to be maximally useful, this has to necessarily be the case, as it helps us to pragmatically conceive of different objects in different contexts to enable understanding.

I will end the discussion with two examples. When we consider a glass of water, we can conceptualize this as one object. We can also conceptualize them as two objects, the glass, and the water contained within. When looking at the oceans we can conceptualize certain areas of them, i.e. the Pacific, the Atlantic, as being individual oceans, and this makes sense as they have differing attributes when considering them at a certain scale. Different species of animal and plant live there, there are unique local conditions etc. If we were to look at drawing specific lines where one becomes the other exactly, we might struggle. At a different scale, we would need to conceptualize all connected bodies of water as one "object", the oceans. Going down to the smaller end of the scale, we can look at water on the molecular scale and the ionized molecules that make it up as well as the molecular minerals and other elements contained in the solution, and to go even further down the scale we can look at the individual particles making up these molecular structures. At this level, there is no difference, and only a few identifiable "objects." There are protons, neutrons, electrons and other fundamental particles. There is no difference between a proton in a molecule of water, and a proton in a molecule of a rock, or a proton inside a molecule of hydrogen inside a star millions of lightyears away. Evoking these larger scale objects to look at the fundamental particles is not useful at this point. Nature does not follow our categorization systems, and categorizations of objects and decisions on what constitutes an individual object at a given scale is necessarily contextual. The oceans do not think or behave as if they are neatly bounded bodies, and the earth does not behave as if the oceans are separate entities from it. These are patterns which we project onto reality in order to be able understand and discuss it. Similarly time in reality is not bounded in the same way that we perceive it. What we consider the year 1886 on a Gregorian calendar for example is an arbitrary viewpoint of a point in time – we do not really believe that the universe had only existed for 1886 years at this point, but there were contextual reasons for categorizing time in this way, just as in general it is pragmatically useful for us to split time into scientific units in order to measure it. We could at any time using a different calendar or different criteria categorize the time period of 1886 or parts of it into other years or beginnings and ends. The universe did not distinguish such a year, or any year in fact, as using the duration during which our planet fully revolves around its star as a measurement of time is something that only we could conceive of. An alien race, even if it measured time as we do, could not conceive of a year in the same way as us unless their planet too happened to take exactly this same period of time to orbit its star. Predictive model theory acknowledges this by including these patterns as part of the predicted model and global model, not the input itself, and stating that the individual decision of what

exactly is interpreted as “event” and “object” is done by humans as needed or in the moment, and can change with context, while “event” and “object” are fundamental priors which our brain uses.

In this Section and Chapter 2 overall I have addressed the second major research question of this thesis: how can predictive model theory help us to understand the systems underlying language? So far, I have argued that we must look at the way we conceptualise objects in our minds in terms of predictive model theory. I have described how the relationship between words and word forms (letters, sounds et.) is shaped by predictive learning and association. I have argued that predictive model theory can elegantly explain the phenomenon of pragmatic normalisation, without needing to evoke multiple contending processes of language parsing, multiple strategies, or contending that humans necessarily use “shallow” processing. In this final section, I have argued that predictive model theory helps to explain the underlying knowledge fields activated, as put in terms of construals of cognitive grammar. In all instances, predictive model theory has helped us to explain, examine, and discuss the phenomena while staying sensitive to the context and individual level of knowledge a speaker or listener might have, and has shown itself to be an excellent tool for analysing language and its use. In the next chapter, I will turn to fictional literature specifically, and discuss the fundamental characteristics of these types of texts which require predictive model theory for analysis.

Chapter 3: Predictive Models in Fictional Texts

3.2 Fictional World Representation: An Ill-Posed Problem

In this chapter I will begin to apply predictive model theory to texts and to answer the third major research question as laid out in Section 1.1: what are the qualities and characteristics of texts which require an approach using predictive model theory in order for us to explain them adequately? This first section will discuss the first particular aspect of the world and texts which creates a unique difficulty for theories of how we process and make sense of the world. I will argue that this difficulty is one that predictive model theory is uniquely able to explain and create hypotheses for. The difficulty pertains to the idea of an ill-posed problem, which is a problem in which the causes behind incoming information is unknown and it is not possible to separate the information into separate effects of different causes. So a theory of cognition and understanding must be able to answer how it is possible to perceive the world and fictional texts within it with only limited information, and how we can understand the world without any mechanism in the world which explains or makes obvious how world processes and mechanisms work.

To begin to answer the questions posed above, in particular the problem of world representation and fictional worlds as ill-posed problems, we must briefly delve into the problems of knowledge and verification. These are not simply problems faced by science or philosophy but by the brain itself. When attempting to reconcile predictions with input and representing a "reality" there remains an open question of how the brain decides what the correct fit is to nullify an error and "explain" an incoming signal. When a fit is found, we are often tempted to say that we "know" what is going on, being said, written down, etc. In terms of a cognitive theory of understanding, we must consider very carefully when we are justified to do so and how to formulate the theory to accommodate issues with this process. As my theory is primarily concerned with texts, the fact that literary texts are fictional must be addressed, and as stated in section 1.1, one of my primary research questions and goals was to analyse the process behind understanding fictional events.

I believe that in order to capture the representational difficulties faced by readers between fictional and non-fictional texts it is clear that a definition of textual fictionality must be able to deal with a variety of phenomena:

1. Clear-cut cases in which the reader encounters a fictional text and knows beforehand that it is fictional, or encounters a non-fictional text and knows beforehand that it is non-fictional
2. Convenient cases in which the reader encounters a fictional text and is easily able to correctly identify that it is fictional, or encounters a non-fictional text and is easily able to correctly identify that it is non-fictional
3. Problematic cases in which a reader encounters a text and is unable to decide if the text is fictional or non-fictional without additional information
4. Highly problematic cases in which a reader encounters a fictional text and mistakenly identifies it as non-fictional, or encounters a non-fictional text and mistakenly identifies it as fictional
5. Hybrid cases in which parts of texts are fictional or non-fictional, and a reader either correctly or incorrectly identifies which are which

Nevertheless, readers generally treat texts as falling broadly into cases 1. and 2. It can be very difficult to deal with the problematic cases, or to give any indication of how or why the problematic cases may come about although they do, sometimes with very serious ramifications as witnessed in current debates around “fake news” and mass distribution of articles and information through the internet with no clear indication of sources or intent but for which it is nevertheless difficult to ascertain if they are true, fictitious, or partly fictitious. For this reason and more we must carefully untangle the notions of knowledge and truth from that of fiction, and in a second step move towards a definition of knowledge and fiction which can better fit the neurological reality of representation as defined by Friston (2005). As I will argue now, our cognitive processes as described by predictive model theory are able to deal with fictional situations in texts because the uncertainty of what is fictional or not, and our inability at times to separate truth from misconception is mirrored in fictional texts by the uncertainty of what is based on the real world and what is fictional. I will begin by describing the difficulty of discovering truth in real processes.

Beginning with truth and knowledge, the most important distinction to be made pertains to the idea of truth values of texts and what it means to have knowledge. For two thousand years it has been held as a quasi-dogmatic rule of philosophy and the sciences that we may have knowledge in the form of “justified true belief.” There is evidence for the beginnings of this definition recorded by Plato in the Theaetetus, in which Socrates considers “that true opinion, combined with definition or rational explanation, is knowledge,” (Plato, 2001, p. 60). In the twentieth century, this view was challenged in a short paper by Gettier who

composed examples of justified true belief which cause great problems for this definition. He supposes that two men, Smith and Jones, are interviewing for a job. Smith is assured by the president of the company that no matter how well he interviews, Jones will get the job. He also knows that by happenstance, Jones has ten coins in his pocket. Because of the direct evidence presented to him, Smith consequently has the justified belief that "The man who will get the job has ten coins in his pocket." To his surprise, Smith gets the job, and, although he was unaware of this at the time, he also later discovers that he had ten coins in his pocket during the interview (Gettier, 1963, p. 122). Smith's assertion that a man with ten coins would get it was right. Smith had very good justification for believing that it would turn out to be right. It is quite clear that Smith nevertheless had no idea he would get the job.

What exactly is the problem? There is an inherent problem in this definition which can be made clear by examining how we define the components of justified true belief. Clearly belief by itself cannot be knowledge because we hold many beliefs that cannot be proven or disproven. They involve ideas about non-observable worlds, concepts such as life after death, and most importantly, new scientific theories. This is where truth, as the objective and irrefutable benchmark of any assertion, is required to ratify the belief. It must be possible to show the truth of the belief by offering evidence from the world of reality. This is unfortunately also insufficient, as it allows for a lucky guess. By randomly making assertions, it is possible to end up being right. On a small scale this phenomenon often occurs in gambling, an activity specifically designed to make it impossible for participants to determine the outcome. Yet people can and do correctly guess it on occasion, or simply observe false causation. The ancient Greek solution which would be accepted until the twentieth century was to add the element of justification to the true belief (Plato, 2001). An individual must also be able to give good reason for the belief, which if true, may then be called knowledge.

The Gettier example makes clear that the element of justification is not sufficient to make the definition work under all circumstances. This situation as described was an ill-posed problem. The knowledge that Jones had better qualifications and the ten coins in his pocket was an interfering piece of information, and we were deprived of the information that Smith also had ten coins in his pocket. This directly affects the way in which justification is defined for this example. Gettier assumes that the correct way to interpret S knowing fact P is that "S is justified in believing that P" (Gettier, 1963). This is based on Chisholm, who interprets the clause as "S has adequate evidence for P," (Chisholm, 1957) and Ayer, who

interprets it as “S has the right to be sure that P is true,” (Ayer, 1956). What exactly is it that gives a justification of a belief however? It cannot be something unknown, and so Smith had to select from the pool of facts he had about Jones. The catch For Smith is that this was an ill-posed problem. The evidence he had for his belief could not be unmixed from the evidence for the final outcome where he receives the job. He was justified in his belief within the context of his own knowledge, but nevertheless wrong because he had interpreted an incorrect causal relationship into the information presented to him.

As readers of the entire story, we do not feel justified in the belief that a man with ten coins got the job by the same evidence as Smith did. Consider the checklist now, and the contents of each clause. The belief is that a man with ten coins in his pocket gets a job. As reader we have been given the information that Smith unknowingly does have ten coins in his pocket. We are also told that he ends up getting the job. For us as readers, it is no longer an ill-posed problem, as we have been given all of the correct causal relationships behind the information. But this is a rare case – most of the time we are Smith, both in life and when reading a text.

A second issue facing us can exist in the form of situations in which more information than strictly necessary exists, and in which an ill-posed problem may lead to an infinite regress. Sanford (2008) gives a perfect example for us to discuss, which is repeated again in Sanford and Emmott (2012), of a person’s causal beliefs about a winter scenario.

> Our knowledge of these particular situations means that we don’t (typically) see the explanation of someone slipping on the sidewalk in winter as being due to a reduction in friction through a thin layer of water forming between a shoe or a tyre and the ice underneath, which leads to aquaplaning. Rather, we see the explanation as slipping on ice because it’s winter. (Sanford, 2008, p. 184)

The example is slightly modified to be about a car skidding on ice, and used to discuss causal attribution in (Sanford & Emmott, 2012). The full explanation for the car is the same as for the person slipping. A layer of water forms on the ice which lowers the friction between a tyre, or a shoe and the surface. The question of truth behind statements such as the above, and of where to draw the line on what is an acceptable explanation, is far more difficult than Gettier’s (1963) example, which at least gave us a definite solution. In order to test whether someone has knowledge about ice, do they need to know how aquaplaning and friction work in order to know that his statement is true? On what evidence do we say yes or no? Does an expert telling us about aquaplaning need to know the physical laws underlying friction, the exact energy exchanges between the surfaces, the forces between the molecular, or atomic particles as they happen, the energy exchanges and forces

between the quarks making up those particles etc.? How can we, in our perception, unmix the causal relationships of all of the individual molecules interacting in the situation and their ultimate effect on the person slipping? Do we need to be made to understand these principles fully ourselves to be able to accept the explanation? It is clear that we cannot ever make a final judgement about where the level of explanation or causal chain behind events actually stops. Instead, we tend to set an appropriate (or at least seemingly appropriate) boundary for explanation, set within the levels of information actually available to us. There may for instance be smaller fundamental particles beyond the quark, but as we have not discovered them, it is difficult to ask meaningful questions about them. They may also not exist. We can however imagine them, and ask questions about these fictional particles, which will be questions about an ill-posed problem as we have no real unmixed causal information to go on. The boundary for all of this is global model integration, whose upper limit is simply the amount of knowledge available to the individual, contained in their global model, as well as the context. As long as a minimal or full predictive model can be formed and made compatible with the global model, the overlap results in a stable neural state and is accepted. The surprising result is that we can accept conflicting truth conditions, by saying that several levels are appropriate. Simple causal statements about slipping on ice, or slipping because it is winter are readily accepted. More detailed explanations are also accepted, when offered, despite the fact that in this case it leads to a conflict. Slipping on ice, the solid aggregate state of water, is not the same as slipping on a surface layer of water in its liquid aggregate state. It depends on how detailed the predictive model was in the first place.

This very same difficulty can occur for fictional texts, and ill-posed problems can be created when attempting to gauge the level of explanation and truth value behind a fictional text. Let us consider a different, classical example of a problem with fictional texts. Searle (1975) believes that in ordinary life there is a systematic relationship between our utterances and their effects on the world, which is not in force in fiction. Following this he asks himself: “how can it be the case in ‘Little Red Riding Hood’ both that ‘red’ means red and yet that the rules correlating ‘red’ with red are not in force?” (Searle, 1975, p. 319). This question is difficult because it contains more questions than it first appears to. Firstly, the information contained in the story of Red Riding Hood is taken in by a reader as indirect input, and in order to even be able to ask the question “Is it true?” we must know what it is we are asking. Like any other knowledge, the verifiability of these facts is also far more complex than can be captured by generalizing whole works as being fictional or not. Searle knows

this and begins to develop a mixed ontology of texts, stating that: “A work of fiction need not consist entirely of, and in general will not consist entirely of, fictional discourse” (Searle, 1975, p. 332). His example is the following:

> Sometimes the author of a fictional story will insert utterances in the story which are not fictional and not part of the story. To take a famous example, Tolstoy begins Anna Karenina with the sentence “Happy families are all happy in the same way, unhappy families unhappy in their separate, different ways.” That, I take it, is not a fictional but a serious utterance. It is a genuine assertion. It is part of the novel but not part of the fictional story. (Searle, 1975: pp. 331–332)

The principle Searle acknowledges is important, but it is very difficult to justify conclusions about the actual fictional or factual status of text extracts. The above sentence from Anna Karenina might well be an assertion, but it is not a factual sentence at all. One might ask what proof of verification Tolstoy had for such a generalization, what precisely distinguishes a happy family from an unhappy one, and how happy families might be happy in the very same way. It is another example of an ill-posed problem leading to an infinite regress. It would also seem very strange to arbitrarily consider any sentences which might have some factual correlation as simply not part of the story, and assume that authors sometimes add sentences that are not related to what they are writing. Authors may well be aware of the mixed nature of their discourse, and are often quite open about drawing inspiration from reality, and attempting to make stories closer to factual truths to be more relatable. It is also problematic from the standpoint of learning. If we are learning about a new topic from a textbook in school or university, we do not know what rules correlate the new information to our experience yet, but we must be able to accommodate this knowledge. If I am learning about the classification of stars and emission spectra for instance, I have no useful prior knowledge, or systemic relationship between my senses to draw on, as I cannot measure wavelengths of visible light with my eyes, see infrared or ultraviolet radiation, or receive radio signals with my brain. Learning about these concepts is no more related to my personal experience than fiction, yet it is possible for me to accept and learn these facts and associate them to the world going forward.

A slightly more inclusive viewpoint which can bridge this gap is made by Currie (1985), while defending a “make-believe” theory of fictional statement. He concludes that a reader may be given to view factual information within a narrative, such as a statement about a real location, with the same attitude as non-factual aspects. This leads him to claim that: “A statement may be both fictional and common knowledge” (Currie, 1985, p. 391). In order for this to be possible, it must be possible for us to conceive of the very same statement without a final judgement as to its veracity, and then to decide in the moment, within the

context the statement is made in, what its status is regarding truth or fiction. I believe this to be the case. In the same sense, this statement also contains a second wisdom: something we know from a fictional story can be both fictional in the sense that it never happened, and real world knowledge in the sense that somebody in the real world really wrote it and produced it in a way that we could read. Many millions of fans around the world agree that in a fictional universe, Luke Skywalker destroyed the Death Star, yet they are not suffering from any mass delusion that this was a real event to be covered by newspapers. This knowledge is not necessarily stored somehow separately, or in different sections of the brain dealing with "fictional" and "non-fictional" knowledge but intermingles as both are activated to deal with specific situations. A distinction of the whole text as fictional or not is often made and kept by an individual as the context requires it, while all kinds of knowledge can and will be used in order to process textual input. This is exactly the solution I would like to propose.

I believe the correct account of the nature of fiction in literature is to acknowledge Currie's point that any given piece of text may be both fictional and not, depending on the viewpoint taken, and that in fact two sets of distinctions are needed. As a first distinction, texts may factually be fictional, non-fictional or contain a mixture of the two elements or even be both. This may include problematic cases in which an author has unknowingly written something factual while intending to write fiction which has either already happened or may come to happen in the future (the reason for the usual disclaimer in novels and movies) or unknowingly written something false while believing it to be fact, perhaps with very good reason. The second distinction must be made within the mind of the reader of said text. Readers will treat texts as either entirely fictional, or entirely factual, or as a mixture, independently of whether or not the text objectively is. That is, when a reader forms a full predictive model of the current text, it will be represented as being a fictional text, or a non-fictional text. During very specific reading situations, such as evaluating a particular research article, or a questionable news source, more fine-grained processing can undoubtedly be used by a reader in order to sift through a text and find erroneous or fictional passages. In essence, the context will decide the reading strategy and thus the level of explanation that will be accepted, which for most normal reading encounters will involve the most efficient processing route, assuming either full fictionality or non-fictionality. There are already differences between reading strategies when the same texts are perceived to be fictional or non-fictional, as evidenced for example by Zwaan (1994), who found distinct differences in individuals' reading behaviour when

different groups were told explicitly that the same portion of text was a news article, or a piece of fiction.

In order to properly analyse the status of given texts as researchers, we must first decide whether we wish to ascertain the precise, objective status of the texts truth values and fictionality, or whether we wish to know what a reader thinks the truth value of the text is. In order to discover the objective status, we can adopt Searle's (1975) strategy and assume that any given text may include both fictional and factual constituents. This can be done not only by looking at entire sentences and their communicative acts, but also at which constituents of a situation are in fact fictional. In Searle's (1975) example above, it would seem that "families" simply refers to the factual concept of families, while the fact which is asserted, that all happy families are alike, is fictional because it cannot be verified. The same, more fine-grained analysis also easily deals with Searle's first example. Red Riding Hood is a fictional character, but this does not mean that the rules governing the reference of the word "red" are somehow not in effect; red is indeed red in the fiction unless explicitly stated otherwise by the author. A girl is also still a girl unless stated otherwise by the author, and so on. Searle was only able to ask this question in the first place because he was able to form a predictive model about the text which he could then compare to his global model, realizing that he only had one definition of red. If there were a fictional meaning of "red" and a fictional causal relationship between the word and a different fictional colour in a fictional world which was not contained in his global model, then he would not be able to consider it in the first place. All words refer to the mental representation an individual has of them, and this representation either has an equivalent in reality or not, but both the rules of usage and reference remain the same. Without knowing it, Searle also used the exact counterfactual and predictive reasoning which defeats the problem: having established in his predictive model that there absolutely must be a difference between word meanings in fictional and non-fictional contexts, it was impossible to integrate them, leading to the quest for a new predictive model that could bridge the gap.

In order to discover what a reader assumes about a text, we have very little recourse except to ask them, or to make predictions. Some experiments such as Zwaan's (1994) showed clear differences in reading strategies when readers were told that texts were of a fictional or non-fictional genre. It is also a general strategy of publishers of fictional texts to shirk responsibilities by placing all published material firmly within fictional boundaries, claiming all similarity to reality to have been accidental, with very little way of verifying if

this was indeed the case or not. For this reason it is easy to presume that most fictional texts, especially those labelled as novels, general fiction, science fiction etc. will be taken as fictional by default. I do believe that readers nevertheless can see the similarities to reality, and always compare such fictional texts back to reality, because the global model of their brain is an attempt to echo reality, albeit from a subjective standpoint. The difficulty in this is that readers know much of what they know about reality from other texts, which they presume to be non-fictional: textbooks, news sources, and websites, as well as taking opinions and viewpoints from fiction. The result is that knowledge about the world and knowledge about texts all integrate into the global model, are remembered by the brain, and are activated in all contexts simply as "knowledge," with knowledge in this case simply being something experienced previously which an individual believes to be true.

The aim of this discussion is to make clear that as mentioned above, clearly delineated boundaries of knowledge between strictly fictional and factual material cannot be easily maintained and the causal relationships behind the input received from a text cannot easily be unmixed in terms of real world or fictional causal relationships. It is rather a question of context and decision in the moment, and readers will over time become rather good at spotting fictional texts and how they tend to differ from factual ones. Knowledge in the first place cannot be fully defined and may always be falsified in the future. Even in cases where we have access to thoroughly researched insight there remains the issue of explanation leading to an infinite regress of further explanations. It is not clear where to draw boundaries and any boundaries drawn are arbitrary and usually the result of reaching the limits of our explanations, or of a need to escape a regress. What a reader knows, the combination of prior knowledge and the input gained from the text itself, defines what can be represented. *A definition of fictional texts and cognition which defines fiction as not true, thus leading to the entailment "non-fiction is true," and that on top of this readers have accurate knowledge about which is which, cannot possibly work or ever hope to accurately capture the epistemic and truth functionality of actual mental representations.* The crux of how our brain deals with the world lies in making predictions about the causes underlying our conceptual experience and updating these predictions as new input comes in. Predictions are "correct" for our purposes when they deliver a pragmatically useful result in recognizing objects and performing actions. Whether predictions correspond to the world as it factually exists cannot be decided from within the brain itself – this must be ascertained from without, which is why humanity requires the scientific method. The scientific method is a third party trial and error system for testing our predictions, and

there is no small amount of irony in the fact that we as a species have copied this exact method of research from our own brain without being aware of it. This same circumstance does not represent a problem however; it is in fact the mechanism through which we learn, extrapolate and understand the complex frameworks of our material and social reality.

Our conclusion must be to state that the inputs and worlds presented by fictional texts present a reader with an ill-posed problem, just as reality does, and a reader must utilise the same predictive models and strategies of prediction and error suppression as they would at any time, even if the final judgement they arrive at is that the text is entirely fictional, or if they knew this already. Predictive model theory allows us to analyse this process correctly, and show how readers can flexibly view texts as uniformly fictional or factual, but when pressed or when necessary to do also consider the same text from a more fine-grained perspective and identify individual aspects of fictionality and realism within it. It also shows nicely how it can be true that fiction can help us to develop social skills and real awareness of our world and how we interact with others, despite being fiction. As will become clear, when learning predictive models as priors for future processing we do not need to clearly distinguish fact and fiction, rather deciding this online as necessitated by the future processing context, allows us to be more flexible in thought, to truly learn from a text even when we are uncertain of its truth, and to glean meaning about objective reality even from fiction.

In this section I have argued that the world and fictional worlds are represented to us as an ill-posed problem. The inputs which we receive through our sense contain information about many causes and ongoing processes, both realistic and fictional and it is often not possible to unmix these. I have discussed how predictive models can as a strategy still lead to outcomes which allow us to pragmatically operate with the inputs we receive, and successfully interact with the world. I have also argued about the limitations of what we call knowledge, and the inherent difficulties of justifying beliefs due to ill-posed problems, and in claiming that a given assertion is also objectively true. I have discussed the nature of fictional texts as also being ill-posed problems, in which it is not always possible to unmix fictional and factual elements and causal relationships of a narrative in the ways it represents the world behind a text, and how again predictive models can help us to draw the line at a suitable level of explanation which can be used to process these. As a final entailment of this I have argued that the way we view information and knowledge as either true and factual or fictional is a contextual decision, which we are not always justified in

making when all related facts are available. We however make the decision based on what is available at the time.

In the next two sections I will build upon this basis of viewing fictional worlds as ill-posed problems by delving into the concept of causality within fictional texts, and the way in which we build up causal chains within our predictive models in order to interpret the world around us, before turning to textual gaps and the ways in which we fill in some of the gaps left by the ill-posed problems, or simply by the underdetermination of texts using causal inferences and prior knowledge. In the next section I will begin by discussing causality and how we construct causality as a concept within our perceptions and representation of the world.

3.3 Creating a Causal Framework of the World

In this section I will discuss the next major reason for adopting predictive model theory for an explanation of how we understand texts and fictionality. This reason is our particular way of representing and interpreting causality and causal chains. I will introduce the particular way in which I believe we view and conceptualise laws and rules, and the minimal predictive model and the basic way that we interpret the world. I will discuss the theory of Barasalou's (2003, 2009) perceptual symbols once more, and then turn to philosophy to introduce Lewis (1973) and the idea of possible worlds. Possible worlds will be used to show how principals which are in essence predictive models have already been used for many years to conceptualise causality. Finally, I will turn to texts and textuality by considering the principle of minimal departure of Ryan (1991), a key principle for textual linguistics, and for my theory as well. This will further show how predictive model theory can fruitfully explain the way readers use knowledge from their global model as well as causal cues from a text to resolve processing of situations they have not experienced before, or which may be impossible in the real world.

The key insight of predictive coding is the fact that the human brain seeks to explain the signals of empirical input it is receiving at all times, by utilizing causal chains beginning with the world and ending at a mental representation. In order to begin constructing how the complex causal body that we call the world is manifested in a brain, we must consider how causality is understood, dealt with, and discovered. Ultimately, the specific causal "rules"

we discover, or think to have discovered, come to stand for “the world” in its entirety. This is the first complication, which we must address before even beginning. Our relationship to rules is a complicated one, ranging from a complete unawareness of rules in most everyday examples of motor control, speech comprehension and production and general biological function regarding our own bodies, to an extreme knowledge of and application of rules in things we consider academic or scientific. The natural sciences often state the primary concern of research to be in discovering the “laws of nature.” It is important to recognize that there is inherently no difference between how we cognitively approach either everyday actions or scientific actions. The major difference is in how overtly we seek to consider such rules, and whether there is a need to store them as separate mental representations.

For instance the brain does not actually contain the rules of how the body works. What it contains is a complete and ever-improving matrix of cause and effect pairings between signals sent to the body and resulting actions without necessarily containing any direct knowledge of why a certain set of motor instructions is more successful than others. This is the same conundrum scientific inquiry faces: We may test and monitor events happening in a linear and temporal order, and we must impose cause and effect onto them without having a clear indication of why some cause and effect pairs work and others do not; this is why human definitions of accepted physical laws simply state that such-and-such causal pairs are always valid. This is the structure of our knowledge about everything, and it nicely shows us how we see the world: the world is a set of causal rules which we expect all objects we encounter to follow. The overall construct is the global model. The immediate question is how learning about the world is structured, or more specifically: how are causal chains identified and structured by the brain?

An initial answer to this question lies in the fact the world is seen as a collection of rules and objects. This is echoed at the neurological level the by the structural ties of forward and backward connections. Forward neurons represent input, while backward connections represent objects and rules and attempt to match them up to input. The most basic causal attribution is the very nature of object recognition. In seeing a chair for example, the basic neurological action is the assumption that the empirical visual input is caused by a chair existing. *Existence is the primordial causal relationship*. This is how any signal is “explained” via backward connections as a recognition. An object is recognized by the activation of already established synaptic connections which match up to the backward connections

necessary. In a case of easy recognition, backward connections likely will not need to modify anything.

The causality in this case is then a simple function of matching. There is not any definable neuron or brain area which is responsible for causality representations: existential causality is our representational understanding of a prediction matching a signal. This matching forms the bedrock of the global model representing the world. We assume the world exists, and within it objects, and that the world itself is in a causal relationship between itself and our senses in which the world causes our sensations.

This means in basic terms that we are naturally inclined to treat forward connection streams as effects and backward connection predictions as causes. This is what is evidenced by Friston's definition of predictions as causes also, but causes are simply taken at face value and causality is not further defined here (Friston, 2005). In order to better represent and explain the causal representations as they arise in the brain utilizing predictive coding, it is necessary to define causality in different ways, and to make clear that causality as the brain represents it is neither necessarily equal to actual objective causality nor to the true nature of causal objects. The second factor here is a limitation of our encoding system. In a binary system such as our brain, the neurological states must be classified by signal or no signal, either 1 or 0. Real causal objects are always both causes and effects simultaneously and it is impossible to empirically observe all possible cause and effect relationships as they happen. Some kind of boundary is needed which can delimit processing of input to be more manageable, and this is achieved through hierarchical predictive processing.

In processing of input, basic causality is taken as above: empirical signals are experienced because objects exist. This is all well and good, but only describes the initial stage and a usually quite unconscious assumption. In the next stage, the brain must attempt to recognize what exactly the input it is receiving means, i.e. what is the object that is being seen, the sound heard etc. If a predictive model is already in place, because a very similar input has already been experienced repeatedly in the past and stored in long-term memory, this is activated and determines what the brain considers to be the cause, based on the input. The causal relationship is now two-fold: first, the signal is perceived because there is an object. Second, the object can be identified because previous patterns of input can match the current signal being perceived, meaning that the object is recognized because it is similar to another object experienced. This rather simple process becomes exponentially more complex as we go further up the hierarchy and combine more signals,

but the basic attribution of causality between signal and processing stays the same. The very same hierarchy governing this system also makes it clear why we perceive causal chains as chains in the first place.

As the complexity of inputs increases, the brain must decide how the various signals and parts of signals fit together into a coherent representation of what is being perceived. The world usually represents an ill-posed problem in which objects are obscured, certain information is missing, or certain circumstances simply cannot be perceived because we have no sensory organs for them. Nevertheless, the brain must attempt to reach a representation that is useful in interacting with the environment. It does so by combining the signals along a hierarchy, beginning with basic object recognition. The next step is to see whether various parts of the empirical signals fit together. As this processing requires more of the brain, it reaches new steps in the hierarchy, which must use the conclusions already reached as the stepping stone for new predictions about what is being perceived. This leads to the conclusion reached at each level becoming the prior for the next higher level, but it also means that at every level a new set of causes must be assumed, which can explain the lower level (Friston, 2005). In other words, there is a relationship of prediction *because* of input between every level of processing, naturally forming a chain of causal relationships as parts of a perceived scene or action are split up into causal chains. Let us consider this alongside an example given by Barsalou (1999) which works very nicely.

Barsalou imagines perceiving a scene of an airplane flying in the sky, which can be analysed in terms of perceptual symbols (Barsalou, 1999). In the scene, initial processing tells the observer that there is an airplane, a large cloud, and some sky. This could happen in stages, but it is equally likely to be a single process as recognition of a plane brings with it the background knowledge that any blue space with clouds beyond it will be sky, or vice versa that some kind of vehicle against background of a sky is likely to be a plane. It does not matter whether there is a slight order to the recognition or if this happens at once. Beyond the object recognition, the brain now has to take some additional steps to fully resolve the scene. The relative location of all the objects takes up higher processing steps, as this must be done after the initial objects have been recognized. Planes and clouds are relatively straightforward, but the sky itself as a background does not give easy indications of cardinal directions. We usually rely on easily identifiable orientations of objects in order to know where up and down is, and of course in a trivial sense in everyday situations we of course feel the effects of gravity on our bodies telling us beyond any doubt which way is down. This very same easy and obvious mechanism does not work when perceiving a scene from

which we have no such feedback, necessitating judgements of relative positions. The locations are compared to the centre of vision of the current observer as well as to each other, and are now also represented as causal conclusions: the plane is left of the centre, and left of the cloud. This is because it is located in a certain area of the field of vision currently being processed. The cloud is seen as being right of the plane, because the plane is to its left. The plane is flying because it is perceived against the sky, with no contact to the ground. The actual interpretation of the scene as an airplane flying high up in the sky and to the left of a cloud is a result of multiple stages of processing, during which causal chains have been formed about what the objects are, what their relative positions to each other and to the observer are, and what they are doing: the plane is flying, the cloud is floating. Because the causal chains recur across hierarchical steps, each higher step not only gives the cause for lower steps, i.e. the plane must be flying because it is next to a cloud in the sky, but must also stay consistent with it and with the global model, which would contain the information that planes fly in the sky, clouds float and so on. The conclusions that it is a plane and the other object is a cloud cannot be reversed at this stage without starting the chain over and each step relies on all previous steps still remaining consistent with global model integration, and in this way hierarchical processing naturally maintains causal chains across representations. It can be nicely shown that through this principle of maintaining causal chains used by the brain, we intuitively attempt to explain counterfactuals, causal statements and so-called possible worlds.

The principal discussion of objective causality, or factually true causality, has been dominated by the analysis of counterfactual assertions since the early twentieth century, although the roots of the discussion reach back some two thousand years to ancient Greek philosophy. The modern discussion has been advanced by addressing counterfactuals and causation through the logic of possible worlds. One of the most famous proponents is David Lewis, who suggests that we may analyse counterfactuals as “statements about possible alternatives to the actual situation, somewhat vaguely specified, in which the actual laws may or may not remain intact,” (Lewis, 1973). Furthermore, the cause within each counterfactual is defined through the presence or absence of the effect or “that if the cause had not been, the effect never had existed,” (Lewis, 1973, p. 557). These alternate situations may then be called “possible worlds” and form a point of reference for the counterfactual statement. It is within these differing possible worlds that changes to the actual situation have occurred. Based upon this, he proposes an ordering between possible worlds upon which the counterfactual is tested, leading to the final definition that “a

counterfactual is nonvacuously true iff it takes less of a departure from actuality to make the consequent true along with the antecedent than it does to make the antecedent true without the consequent," (Lewis, 1973, p. 560)(iff here means "if and only if," a common phrasing of stressed necessity in philosophical writing).

Several scholars since have argued that one should interpret possible worlds as various sorts of mental and even textual entities (Menzies & Pettit, 1994; Naylor, 1986; Rescher, 1999; Stalnaker, 1976). I agree with this view and also assume that possible worlds are in fact mental representations used for the purpose of reasoning. Each considered possible world is a full predictive model, which must be compared to the global model representing our knowledge. The one which is most similar is chosen as comparison, and if the counterfactual can be integrated to both, it is considered true. In the hierarchy, the argument must always begin with considering the precedent, the "if x were(not) the case", followed by the antedecent "then y would(not) be the case" and finally, whether this fits with the world considered. The antecedent cannot be considered true if the precedent is voided by the consideration and causes further errors, or if it is impossible to imagine a world in which the precedent cannot occur. Similarly, the antecedent cannot be true within the hierarchy if it is not accepted that it would follow from the precedent, that is, if the precedent does not move up the hierarchy as a higher-level cause for the antecedent. This then happens naturally, and nicely shows how we make such considerations effortlessly: if it all works out, the precedent is considered and represented as a minimal predictive model, followed by a representation of the antecedented having happened or not happened because of the precedent which is now a part of the prediction moving forward. If the antecedent is consistent with the prediction based on the precedent, the precedent becomes cause for the new input, and moves forward to be integrated with the global model. When the integration to the global model is accepted, the precedent does not suddenly become causal, but in the process of getting there it already was considered a cause. Similarly, if the integration fails, it is because it was already rejected as having a causal relationship to the antecedent, by virtue of a failure of the prediction. However, in order for us to ever be able to learn something new or change our global model, there must also be some flexibility in integration, and this flexibility is nicely described by Lewis' (1973) model of similarity: that is, it may not always be possible to have a perfect fit of integration with the global model, but it may also be undesirable to outright reject a predictive model we have constructed. In this case it seems useful to suppose that we will select the closest possible predictive model to our global model and accept it, overriding any last remainder

of error. This allows a spectrum of potential best fits for integration with contextual criteria for when a model can be accepted or not accepted.

This precise set of reasoning has been imported into theories of literary interpretation such as Text World Theory which make use of possible world theory (Gavins, 2007; Ryan, 1991; Werth, 1999). Ryan is one of the first to bring possible worlds into the study of literature, and to recognize that the reading process must be akin to counterfactual reasoning in the sense that just as we suppose alternate worlds to reality for causal reasoning, we can suppose the worlds described in literature to be alternate worlds in a very similar sense to Lewis' description. These alternate worlds are signalled by "world-creating predicates", which are formulations of counterfactual sentences, verbs such as to wish, to dream, to plan or others (Ryan, 1991, pp. 19–20). A reader confronted with a world-creating predicate creates and steps into an alternate world, which according to Ryan is considered the actual world for the duration of the reading process, leading to more possible worlds in the forms of the wishes and considerations of characters within the actual world of the text (Ryan, 1991). This process is always incomplete, as no human has the mental capacity to fully represent every possible property and detail of the object or objects in question (Ryan, 1991). Not only is the description of objects incomplete in this way, but also the description of these worlds. In order to complete those parts of the possible worlds considered, Ryan suggests a very important law of not only possible worlds, but as I see it, predictive models:

> This law—to which I shall refer as the principle of minimal departure—states that we reconstrue the central world of a textual universe in the same way we reconstrue the alternate possible worlds of nonfactual statements: as conforming as far as possible to our representation of AW.[2] We will project upon these worlds everything we know about reality, and we will make only the adjustments dictated by the text. (Ryan, 1991, p. 51)

These considerations offer good evidence that the cognitive process of reading literature overlaps with the process of understanding counterfactual or causal statements. It also adds another dimension to the process however, since fictional texts are underdetermined. This is used creatively by authors of fiction. As Ryan puts it: "True fiction exploits the informational gaps in our knowledge of reality by filling them in with unverified but credible facts for which the author takes no responsibility (as would be the case in historiography)" (Ryan, 1991, p. 34). Nonfictional texts meanwhile do not depart from reality and strive to the ideal of absolute compatibility between the actual world and the world created by the text (Ryan, 1991). I believe this further underlines very well our intuitive approach to texts,

[2] Ryan's abbreviation for Actual World; the real world we physically inhabit.

and the way we predictively use all of our real world knowledge to process fictional texts, even when treating the entirety of the text as fictional. While the epistemic problem regarding knowledge of how closely a textual world aligns with reality remains, it does appear to be an important part of our understanding of fiction and non-fiction to know that in one case there is no genuine attempt at representing reality or responsibility for accuracy, while in the other case there is.

An evolution of some of Ryan's work and possible worlds theory which enjoys widespread contemporary popularity is Text World Theory, first proposed by Werth (1999) and brought to maturity by Gavins (2007). The theory maintains many of the aspects taken over from possible worlds theory, also using the concepts of "world-building elements," but defining them more closely as consisting of pronouns, locatives, spatial adverbs, verbs of motion and deictic expressions (Gavins, 2007, pp. 36–38). These elements, once reading has begun and a reader has already formed a textual world, go on to drive "world-switches" which cause the reader to represent a new text world as the textual description shifts to different spatial or temporal locations (Gavins, 2007, pp. 48–49). This overlaps well with what has been discussed above, and in fact despite Text World Theory per se using no neuroscientific sources it can quite effortlessly be integrated into predictive model theory: Text Worlds represent amalgamations of high level predictive models in which the changes of time, location or action within the text must be represented and explained with a new set of causes. However, I disagree with the notion that text worlds as described in the theory properly reflect mental representations during online processing, and suggest that they rather reflect idealized representational constructs gained through re-reading and analysis, discussed in more detail in Neurohr (2019). Importantly, this theory suggests that it is usually the text which drives inference, also called the principle of "text-drivenness" (Gavins, 2007, pp. 36–38). In a somewhat different way this states the same thing as what Ryan points out above, namely that the text leads to a filling in of the gaps in background knowledge using fictional material which is integrated by the reader.

What makes this construction of causal relationships within representations so interesting for the understanding of texts lies in the nature of how high level causes are actually generated in predictive models, and how they are selected within a situation. One of the interesting constraints appears to be that whenever high-level causes must be constructed in order to explain the current input, the brain prefers to integrate the information of the

input itself as much as possible. This is evident within Text World Theory for example in the way the inputs of the text are portrayed as driving switches between representations and supplying the information necessary for building the new representations. It is also evident in theories of causal and counterfactual reasoning about the real world as evidenced in possible world theories. Both cases support the supposition that: *whenever part of an input needs to be explained, our brain will preferentially do so using causal factors within the input itself and assuming causal relationships to be present before adding further inference from memory or testing these against retained real world knowledge*.

In this section I have argued that we conceive of objects existing as a fundamental causal relationship. I have discussed how this leads to causal perceptual chains of objects causing our perceptual inputs, which in turn cause explanations to be needed and turning into complex causal explanations. I have discussed how Barsalou's (1999,2003, 2009) perceptual symbols can account for the causal nature of our perception as an effect of objects' existance. I have showed how many of our explanations for other causal statements and states of the world are explained by possible worlds in philosophy, which is a method utilizing what are in essence predictive models. I have then turned to the principle of minimal departure, and the idea that when we have no access to causal information, we utilize any information in the input or text itself, followed by any information in our global model.

The exact order of which causal information precisely is used at any point is constrained by other factors, and distinguishing between all the possible factors is difficult. Let us begin with the assertion that causes from within an input receive high priority nevertheless. There is good reason to believe that this is a very principal operating strategy for our brain, and that this forms a crucial part of reading strategies in all texts. It can be illustrated by considering the phenomenon of textual gaps and underdetermination, which I shall turn to in the next section, where I will discuss the idea of gaps as put forward by Ingarden (1968), but also Iser (1972), Barsalou (1999), and Schank and Abelson (1977). I will then turn to some actual textual examples I have selected to show gaps, and selective causal information which can be used by readers to fill them.

3.4 Textual gaps and causal inferencing

In this section I will address the third important trait of both the world and texts requiring a predictive model theory analysis, which is underdetermination. I will explore the fact that it is at no point possible to have all possible information about a given object or event. This is a distinct point from the contents of section 3.2, as here we are not discussing information in terms of the epistemic difficulties surrounding a text's fictional status. Rather, I will be addressing the strategies by which readers fill in the gaps left by ill-posed problems and missing information within the texts, and how predictive models can fruitfully lead to stable interpretations based on what I have argued so far in sections 3.2 and 3.3. I will begin by briefly discussing the notion of these gaps as brought forward by Ingarden (1968) and Iser (1972).

One of the communicative limitations of literature and the events described by a text, is that by necessity a text cannot mention every possible detail about the proposed events being described. Ingarden calls these gaps and considers such gaps to be far from problematic. According to him, they constitute one of the main pleasurable effects of reading literature. "Gap" here refers to the phenomenon that a reader cannot tell whether an object or a character has a given property or not, because it is simply not mentioned. A reader attempts to use clues contained within the text to postulate what these properties might be and to attempt to fill in the gaps in a way which fit the text and lead to aesthetic pleasure (Ingarden 1968). The pleasure is derived from actively interpreting and engaging with the text as a reader. Iser (1972) builds upon this idea of gaps and considers them a necessity not just for aesthetic effects but for the understanding of texts, which an author utilizes to construct the process of interpretation within the reader. Stories which are intentionally underspecified force a reader to attempt to interpret the missing information in a certain way, in order to make sense of what is happening and to gain a sense of meaning (Iser 1972). Filling in the gaps of stories in this way suggests a similar priority of inference as discussed in the last section, but then faces a new difficulty head on: what happens when the information needed to predict a cause within a given situation is simply not contained within the text itself, or in some cases, not available anywhere?

I will outline a possible system of inferential reasoning, which can nicely be based on predictive model theory. The possibilities of this priority system will rely on the fact that we

form predictive models and learn from them or possibly even learn them in their entirety as forms of schemata/frames in the spirit of Minsky, Schank & Abelson (1977), and van Dijk & Kintsch (1983). The same possibility is also suggested by Barsalou in perceptual symbol theory (1999), and it aligns quite naturally with all the principles I have adapted from Predictive Coding. Whenever certain connections are activated regularly to explain the same input and lead to a stable state this will naturally lead to Hebbian learning and to the entrenchment of the predictive model in memory. Considering the fact that reading literature, whether fictional or not or a mixture, consists of the same processing, many of an individual's stored predictive models and resultant predictions about inputs will come from fiction. However, as I have said above, drawing on other research, the principal of minimal departure seems to apply and dictate that we always prefer to compare situations to reality and our perception of the real world. I would like to suggest that this does not create a problem, or indeed any kind of contradiction so long as we remain careful to look at how our perceptions of both fictional information and reality might look and remember that as discussed so far in sections 3.2 and 3.3 we do not need to know if our predictive models are objectively true, only that they fulfil the criteria of being plausible and forming a satisfactory causal chain. Let us look at some textual examples which require both textual and real world knowledge in order to fill gaps and suppress errors to illustrate this.

In many fictional stories, unexpected occurrences may happen, in which the aspect of the action which is causing surprise is not explicitly described. Consider this passage from a fantasy novel featuring a magical sword wielded by the protagonist:

> Annoyed, Richard took the sword in both hands, feeling the anger surge through him. He gave a mighty swing at the remaining tree. The tip of the blade whistled as it sliced through the air. Just before the blade hit the tree, it simply stopped, as if the very air about it had become too thick to allow it to pass.
> Richard stepped back in surprise. He looked at the sword, and then tried again. Same thing. The tree was untouched. He glared over at Zedd, who stood with his arms folded and a smirk on his face.
> Richard slid the sword back into its scabbard. "All right, what's going on." (Goodkind 1995, p. 126)

What is remarkable about this passage? At face value not much, and it is unlikely that any reader would hesitate over the concept portrayed in the passage, or the surprise that the character feels. As readers we are likely to be just as surprised at the idea that a sword swung by an adult man with both arms at full strength should simply stop, mid-swing, stopped by nothing but air. The passage makes obvious what is otherwise unconsciously

assumed across the entirety of the text, or indeed virtually every text in existence.[3] Newton's laws of motion are by default always in effect and only become noticeable aspects of a story when they are violated. Even in the context of the story however, this violation is an anomaly. The text never explained to the reader that every solid object maintains a constant velocity as long as it is not opposed by other objects (Newton 2011). The fact that this rule is broken and the sword is stopped in mid-motion without colliding with another body nevertheless causes surprise in both the reader and the character of the story. In this case, no genre-specific knowledge was necessary, and as shown in figure 13 this part of the passage elicits an error by violating our real world knowledge of the way objects behave when swung under normal conditions.

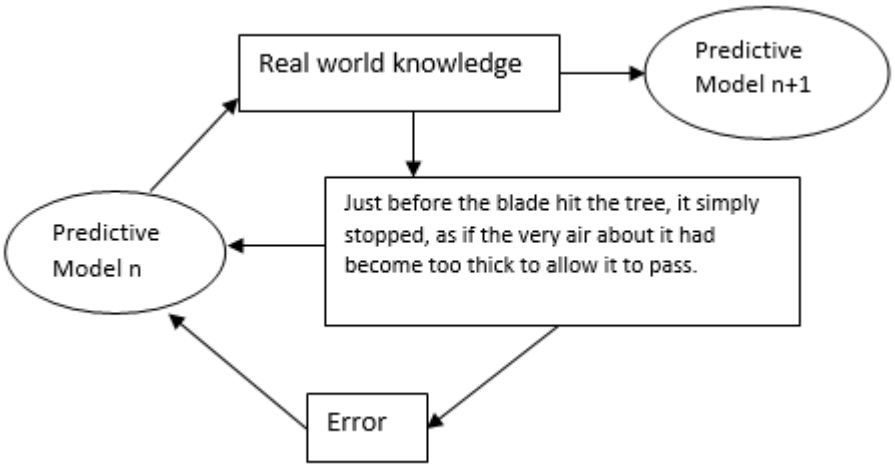

Figure 13. By violating real world knowledge and offering no other causal information, the passage elicits an error within the predictive model n, necessitating a new predictive model n+1

This is an example in which the author knew and assumed that the laws of motion demand a different outcome and frustrated them for the effect of the story. A reader is likely to form a new prediction which contains a yet-to-be-discovered cause for the otherwise inexplicable effect. The completion of this aesthetic subversion makes it necessary for the text to also reconcile the problem, and in the following lines it is explained that a kind of magic causes the sword to behave according to different rules than simple Newtonian

[3] This works even with anomalous texts; there are some works of postmodern literature which actively seek to undermine even such basic constants of logic and our perception of the world. The impact of such work comes from a disruption of the expected physical laws, and the disruption itself becomes an expected trope.

motion (Goodkind 1995). In cases where a text potentially does not offer any useful explanation, it is also possible for the reader's brain to accept whatever prediction it has made regarding the cause of the error, and to cancel out the error without the excess input. This is the same strategy which is behind cases of pragmatic normalization as discussed in sections 2.2 and 2.3, in which an individual can decide to cancel out errors caused by a statement through pure prediction even though the statement itself does not provide the input necessary to do so.

This example makes the usually completely unquestioned fact obvious: the laws of physics and of real world causation are the default assumptions underlying all basic motion and action in stories and they are not explicitly taught or even mentioned in literature unless they must be amended for the sake of fiction. Causal chains themselves always follow this real world knowledge structure which will be contained in the reader's global model, and accordingly the text must explicitly provide an alternate cause for events if the default cause and effect structures of reality are violated. This is a kind of gap which a text creates in almost all descriptions of events, and only fills out through the story if there is a particular need to supply information which goes beyond the default expectations. In most circumstances, assuming the laws of physics to hold is a triviality, much like the omission of the colour of a chair, or describing a particular kind of table. Any reader still assumes that a chair will have some colour, and a table some kind of shape. Similarly, any reader knows that a kicked ball will fly away and that the usual causes have the usual expected effects. For the sake of developing an adequate theory of how meaning is formed, this fact is not trivial at all. In order to fully develop and resolve predictive models for the above examples, a reader by necessity had to apply the principle of minimal departure, but this did not lead to only applying knowledge supposed by the reader to be factual: The principle in this case also covered the application of knowledge the reader knows fully well is genre and story specific, but which forms part of fictional stories which are part of our everyday lives, even though the knowledge itself does not apply to the material world around us.

Not only general or abstract world knowledge is mixed within literary understanding, but also more specific object knowledge can become mixed within predictive models necessary to understand some texts. An interesting mixture of real world knowledge and genre specific knowledge is required by a more detailed explanation of special gemstones in the following extract:

> *Conjoiners*: By infusing a ruby and using methodology that has not been revealed to me (though I have my suspicions), you can create a conjoined pair of gemstones. The process requires splitting the original ruby. The two halves will then create parallel reactions across a distance. […]
> Conservation of force is maintained; for instance, if one is attached to a heavy stone, you will need the same strength to lift the conjoined fabrial that you would need to lift the stone itself. [...]
> *Reversers*: Using an amethyst instead of a ruby also creates conjoined halves of a gemstone, but these two work in creating *opposite* reactions. Raise one, and the other will be pressed downward, for instance. (Sanderson 2014, p. 1085)

A reader must know and understand what the basic types of gemstones are, as Sanderson (2014) only refers to the same typology of gems we know from reality, in this case rubies and amethysts. The creative abstraction is that these gems are given causal power over factual physical forces and mechanics. Nevertheless, the basic expectations of these factual forces remain intact: a force which acts on one object is transferred to another object, or the opposite force is transferred but the overall mechanics of real world forces are observed. Forces pull or push objects in accordance to the direction they are applied in, and the basic effects of the gemstones closely echo the type of effects we can observe from magnets in reality. In order to understand the fabrials he speaks about, a reader needs to know what rubies and amethysts are, and real mechanics of motion and causation to tie the force affecting one half of the respective gem to the other half and in turn understand their relative motion. The final predictive model must combine these concepts together and allow the gemstones themselves to not only be causally affected by forces, but for their halves to causally affect one another.

Other examples, where real world scientific knowledge is blended into a text can however lead to an interesting phenomenon in which a reader may or may not be aware of which portions of textual knowledge are in fact factually accurate or not. Consider this section from Ian M. Banks' (1992) *Use of Weapons*, which talks about a small, highly advanced spaceship designed for virtually undetectable take-off:

> Watching from the ground - if they hadn't blinked at the wrong moment - a very keen-eyed observer might just have seen a column of trembling air flick skyward from the summit of the keep, but would have heard nothing; even in high supersonic the module could move more quietly than any bird, displacing tissue-thin layers of air immediately ahead of it, moving into the vacuum so created, and replacing the gases in the skin-thin space it had left behind; a falling feather produced more turbulence. (Banks 1992, p. 48)

At face value, this situation differs from the other examples discussed so far, as it does not introduce either magic or alternate natural resources. Instead, Banks (1992) introduces advanced technology as a concept for an extremely quiet spacecraft. In truth, this example does precisely the same thing as all the other examples so far. The real-world knowledge which must be activated in order to understand this situation is simply more elaborate, while the fictional additions to it are more subtle. The difference lies in the interpretation of what kind of knowledge is being activated. A reader must first of all conceptualize the "displacement" of layers of air from in front of the craft to the area immediately behind it. How this is achieved is not explained by the section, nor any section within the vicinity of this text sample; the text merely tells the reader that the society which built the ship has "displacer" technology which can move objects from one place to another. This immediate part of the presented causal chain can easily be predicted by textual knowledge given within the text and integrated to cancel out errors at this point, as shown in figure 14.

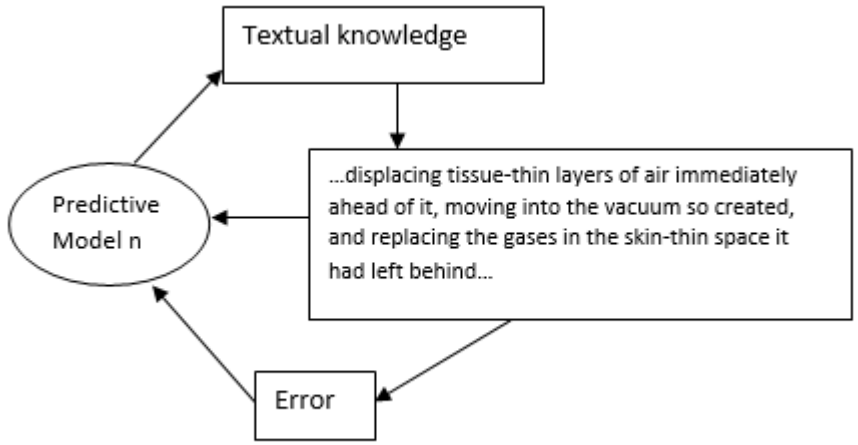

Figure 14. Specific textual knowledge can account for an error elicited by this section.

Taking this fact in stride, a reader must then answer the question as to why displacing air from in front of the craft would actually make it silent. This is not necessarily general knowledge, and comes from actual physics: the noise made by flying aircraft is produced by various sources, one of which is the flow of air across the body of an aircraft, also called turbulence, which in turn causes sound waves to radiate outward from the forward motion of the craft (Hubbard 1991). The actual physics behind this concept is quite complex, and even an educated reader may only be aware of the basic principle that sound waves are produced by a flow of air. This is the only way to actually understand the point made by the text; the novel in its entirety does not contain any treatise of a fictional aerodynamics or

aeroacoustics to account for any departure from real physics. Similarly, the idea of displacing layers of air has no equivalent in reality, but depends on the physical principle that an object, in this case a volume of air, can be moved and that something must occupy the space it has left behind. A reader might well not be aware of the factually accurate concept of turbulence used here, and incorrectly assume the entire section to be fictional and have no equivalence to reality, simply accepting the causal chain presented within the text. This means that two actual processes are possible: the reader can be aware that factually accurate principals of turbulence are mixed into the fictional textual information presented, or the reader can view the entirety of the input as being fictional textual information which simply uses a familiar word. As shown in figure 15, both amount to the exact same process of understanding, merely differing in the internal labelling of whether the reader perceives parts of the information to be accurate of reality, or not.

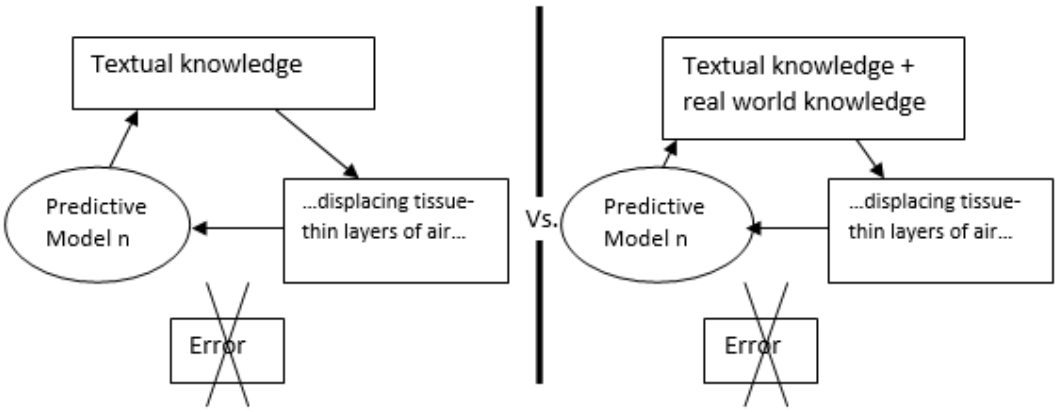

Figure 15. Two possibilities of internally distinguishing the knowledge used to cancel out error resulting from textual input

Compared to the first examples of this section, which only relied on general genre expectations and supplied a generous amount of specific textual information, the above quote relies almost entirely on genre tropes of science fiction. Where the example of Brandon Sanderson is fleshed out with fictional information, offering a full underlying explanation for the phenomena described, the displacer technology of Ian Banks merely serves as a plot device to fictionally accomplish small situations which need to be achieved without grand explanation, in which case causal displacement, like a trope, is briefly mentioned. A similar thing happens later on in the novel, as a character futilely attempts to fire a weapon inside an advanced spaceship. After multiple unsuccessful tries, another character, a sentient machine, suggests asking the ship to make a change in the bay they are occupying: "'Try asking it to clear the bay for firing practice,' it suggested. 'Specifically,

ask it to clear a space in its trapdoor coverage'" (Banks 1992, p. 135). Upon making this request, the attempt to fire again is a success, to the surprise of the character. While exiting the bay, he asks the machine what precisely it meant with trapdoor coverage, and receives one line of explanation: "'General Systems Vehicle internal explosion protection,' the drone explained, […]. 'Snaps anything significantly more powerful than a fart straight into hyperspace; blast, radiation; the lot'" (Banks 1992, pp. 136-137). By this time, a reader will have had occasion to be informed of this technology existing several times throughout the novel, and of course potentially from reading other novels in the series of Ian Bank's Culture novels, which feature a recurring advanced society and fictional universe. The genre expectations of science fiction in general allow for advanced, possibly incomprehensible technology, and in fact demand some presence of it in a text in order for it to qualify as science fiction. The original problem mentioned is yet another simple case of disturbed causality: a gun is supposed to fire, but to the frustration of one character who is not sufficiently familiar with the technologies involved, it will not. A cause is left without an effect and both the character and a reader require the causal chain to be completed somehow. To show how advanced the technology is, the example is framed within the story to make the male human character look uninformed, and the drone to consider it a trivial matter, which only deserves a passing explanation. The causal chain is completed by mentioning that it is possible for the ship to simply 'snap' anything it wants out of its interior straight into hyperspace (another science fiction trope which the author does not attempt to define). The gun was firing all along, but the ship was simply 'snapping' the resulting shot somewhere else to be safely discharged. The problem incurred is solved within a predictive model featuring the knowledge of a kind of technology which exists within the fictional world, a causal chain which needs to be completed, and a final piece of information which completes the causal chain and satisfies the active event schema.

This circumstance, in which either real world physics knowledge or genre knowledge can be sufficient can also happen in other genres, or alongside examples from the same text where it was not the case. Let us return to Goodkind (1996), the author from our very first example, featuring the magical sword. The other character mentioned in the extract, Zedd, is a wizard, who is shown having some interesting talents of his own.

> Instantly, Zedd ignited the air above the water, using the heat in the water to feed it. The wizard's fire sucked all warmth from the water. The entire pool froze into a solid block of ice. The screeling was encased. The fire sputtered out when the heat feeding it was exhausted. There was sudden quiet, except for the moans from the injured across the hall. (Goodkind 1996: 14)

Part of this scenario must necessarily be understood by utilizing specific knowledge of genre introduced both within the novel itself, and other high fantasy stories of a similar kind. The character of Zedd is a wizard, which has cultural implications, as wizards are a stereotypical character archetype which can be found in a huge body of literature, but also specific implications set up by this author. The book itself and its predecessor tell us that wizards are capable of magic, and that Zedd is able to summon fire or lightning at will. Nevertheless, the text-specific knowledge of a wizard coupled with the genre knowledge of what wizards and magic can do are not the only knowledge a reader may use to comprehend the passage above. The knowledge contained in the novel, together with genre expectations of magic and how wizards function in stereotypical stories leads to a predictive model describing that the wizard was able to make fire. The way in which this is done can also be understood in terms of real world energy transfers if a reader has knowledge of these. It encompasses the laws of thermodynamics. It is also brought specifically to the attention of the reader as this section, while being underdetermined on many aspects, goes into the detail of mentioning that the heat is drawn from the water. Generic magic in such stories would easily be able to simply assert that a wizard can create the heat through magic itself with no further effort. This can be graphically represented as in figure 16 below, assuming that a reader comes to this passage with a given prior predictive model n, and now encounters this passage of text. As shown, predictive model n coupled with previous genre knowledge of the nature of a wizard is likely to produce a prediction which can match the first line of the extract well, and integrate the representation of a wizard igniting a flame. When the second line is input however, the background knowledge specific to what a wizard is, will not be sufficient to integrate the concept of a fire having “sucked all warmth from the water” and will lead to an error which must be resolved as in figure 16 below.

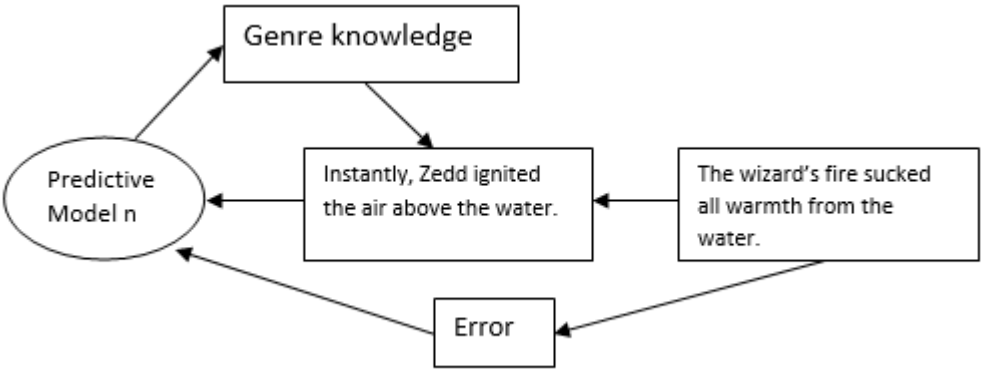

Figure 16. Genre-specific background knowledge can be used to successfully predict the first line of input, but not the second line, leading to an error feeding back into original predictive model *n*.

Readers now have recourse to two strategies, as they did in Banks' (1992) example of advanced technology. We can understand this example from the point of view of genre tropes entirely, saying that wizards would have the ability to manipulate heat, transfer it around and create flame as they wish, and fulfil a genre plausibility. Alternatively, if we a reader has the background knowledge, they can recognize this as an application of real world thermodynamics, and the law of conservation of energy: "Energy can be transferred from one form to another, but it cannot be created or destroyed. The total amount of energy always stays the same" (Johnson et al. 2000, p. 64).[4] This very same law is also the first law of thermodynamics, stating that the change in internal energy of a system is given by the sum of heat transferred and work done. This law interacts with the so-called "Zero-th" law of thermodynamics: "If two systems are at the same temperature, there is no resultant flow of heat between them" (Johnson et al. 2000, p. 298). The author of the above passage, may have done so accidentally, or he has learned these laws either implicitly or explicitly through school and personal experience. We know that heat transfers between objects and we also know that if water has sufficient heat removed from it, it becomes ice. Conventional knowledge, with no explicit rules which were learned through instruction, would likely lead to the generalization that ice is water which is "cooled." From the standpoint of thermodynamics, this is incorrect and the author has correctly applied scientific knowledge: heat is transferred from the water into the air above it, because the temperature equilibrium between them is changed. Just that in this case, the cause of the

[4] This citation and the other physics examples are chosen from the textbook I myself used while completing physics A-Levels in school; like many thousands of other students taking physics.

heat transfer was a wizard. By combining the genre knowledge which could be used to predict and resolve the first line of the extract with this knowledge of factual thermodynamics, it is possible to form a prediction which can resolve errors from the new input and lead to a new, stable predictive model as represented in figure 17. While the genre knowledge accounts for the nature of what "wizard's fire" is, real world knowledge of thermodynamics combines with it in order to explain the transfer of heat from water to flame. The same thermodynamics knowledge now accounts for the entirely realistic process within the next line, as the water freezes due to the sudden loss of all heat. The resultant predictive model n+1 contains the mixture of knowledge used to cancel out the error and the representation that this specific instance of "wizard's fire" behaves in this specific way and is able to freeze water as in figure 17 below.

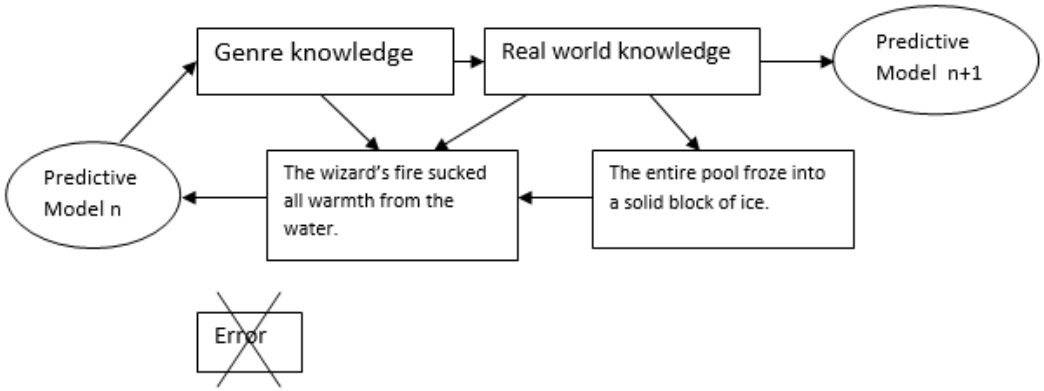

Figure 17. Combining genre knowledge with real world knowledge of a physical law allows for a cancellation of error, and integration of input into a new stable predictive model, n+1.

The end result is similar to the example from Ian M. Banks (1992) above, and readers can choose to rely entirely on genre expectations and the textual information, or they can choose to fill in their own background knowledge of thermodynamics, as in Figure 18.

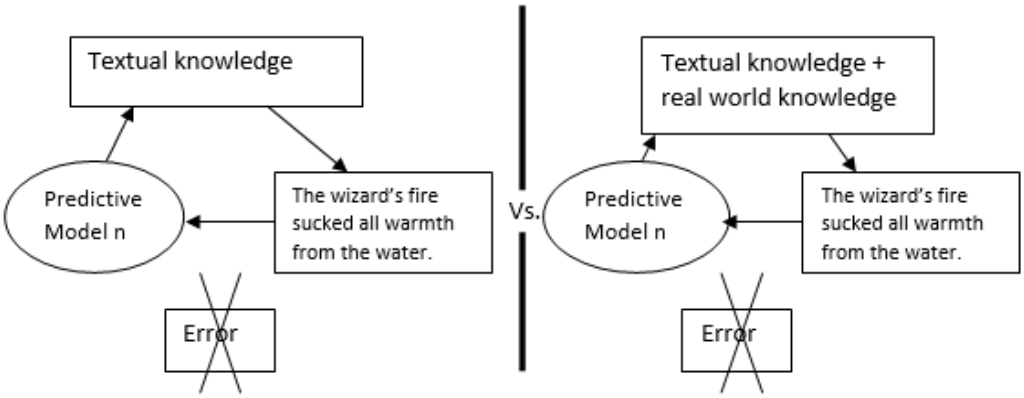

Figure 18. Two possibilities of internally distinguishing the knowledge used to cancel out error resulting from textual input

As can be seen from all of these examples, the most important part of understanding any textual sections via predictive models is for the situation to fulfil the one-to-one principle I have outlined in Section 1.4, and to represent a completed causal chain without violations of time or space. In order to complete the causal chain satisfactorily, any mixture of fictional or non-fictional knowledge can be recruited, and the model can satisfactorily explain a given input even with very little detail. The fact remains that real world knowledge preferentially fills textual gaps, but it can be freely amended by fictional knowledge both from within the text itself, or from specific genre expectations stored as part of a reader's global model. Just as fictional knowledge is mixed up with real world knowledge, recruiting real world reasoning and logical inferences, this relationship can work in the opposite way as we are often tempted to recruit fictional reasoning and examples into considerations about reality.

In this section I have argued that texts, much as reality, have gaps, which are occurrences of missing information about a specific object or event. As finite things, texts only contain a limited amount of information, just as our senses can only perceive so much about any given object at a time. I have argued that the principal strategy for overcoming these and filling in the gaps appears to be that of utilizing predictive models as I have described them. I have shown how explanations using predictive models can account for reader's interpretations of multiple real textual examples, and even how some examples can be resolved independently of the level of individual real world knowledge of readers by following the one-to-one principle as well as the principle of minimal departure.

In this chapter as whole I have so far demonstrated the following:

Predictive model theory is a context-sensitive theory of textual understanding with great explanatory power. I have argued for the validity of using this approach by showing that three specific conditions are true about both texts and the real world, which in turn answers our third research question of the thesis: What are the qualities and characteristics of texts which require an approach using predictive model theory in order for us to explain them adequately? These qualities are the following:

1. The world and texts are ill-posed problems, in which we cannot resolve the source or cause of the input we are receiving or un-mix multiple causes which are present, and we must attempt to do so without outside guidance by relying on our prior knowledge and reasoning capacities. We use prediction to posit the existence of the causes which may be responsible, and test our inputs against these until finding an acceptable result. This may not always be the correct result in terms of the reality we live in, but our process will terminate when we have achieved a level which allows us to successfully interact with the world based on our conclusions. “Good enough” in terms of our understanding, is generally equivalent to “correct” unless we have a specific reason to delve further into an explanation.

2. The world and texts are represented by us as complex webs of causal interactions, and we are always attempting to find the explanation as to why a particular event is occurring, and to predict the next step. We understand both the world and texts in terms of the underlying causal laws and principles which we believe to guide them, and these form our expectations of how objects and people will behave.

3. The world and texts exhibit gaps, because they are underdetermined. Texts specifically are underdetermined because they do not contain all information possible, while the world is represented to us as underdetermined because our senses are not sufficient to take in and process all information about it. Therefore, we do not at any given time have access to all of the information about a given situation, and we must fill the gaps through prediction and the use of global model knowledge from past situations. Many fictional texts rely on the global model knowledge of readers, only modifying the parts necessary for a plot point or situation, and readers will follow these text-internal clues before utilizing their global model for any remaining gaps by forming predictive models.

The next step is to consider how these characteristics are solved by predictive models in a real time reading process. This was also tested by myself using an empirical experiment. In the next chapter, I will introduce and discuss the experiment as well as the principle I will

call contextual plausibility, and the consequences of the results for my theory and my future research.

Chapter 4: An Empirical Test of Predictive Models

4.1 Experiment: Predictive model dimensions in natural text

In this chapter I will discuss the first of two eye-tracking experiments which I performed as part of my research into predictive models. I shall introduce the thinking which lead to the experiment, including my own planning and intentions in Section 4.1 and then present the experimental conditions and results in Section 4.2. The remainder of the chapter, Sections 4.3 and 4.4, will be dedicated to discussing the results and then offering the next stage of the theory which sprang from the results directly and led to further research of mine. Together with chapter 5, this chapter will begin to answer the final and most important research question of this thesis: using predictive model theory, what does it mean to understand a fictional text describing events which never happened and how does this happen in a typical reading process? (see Section 1.1).

The first experiment was intended to investigate how predictive models could be used in an online reading process. Specifically, I was interested to discover if the one-to-one principle I proposed was visible during natural reading and if there was any measure of priority between causal, temporal and spatial representations and plausibility for readers. By priority I mean here if inconsistencies in one of the particular dimensions causes a greater disruption than inconsistencies in others, which would in turn allow me to conclude that it is more important for successful processing than any others. In particular the goal was to add to the results gained by previous experiments designed to test the situation model dimensions (or in this case predictive model dimensions) of the "event-indexing model" (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998) by using a whole fictional text and state of the art technology which can capture a real-time reading process in some approximation to natural reading. The event-indexing model experiment considered five such dimensions: time, space, causation, motivation and protagonist. They made two global predictions based on the hypothesis: 1) that the brain will more strongly associate events which closely overlap on these five dimensions, i.e. sharing a protagonist or location, and that 2) events which do not overlap are more difficult to associate to another and the more cognitive "effort" must be expended in order to construct a situation model containing both (Zwaan, Radvansky, Hilliard, & Curiel, 1998, p. 200-201). The researchers adapted well known fables and rated the connectedness between events within the texts. They were then read by participants in

a self-paced reading task, which recorded the reading times. Their results were surprising to them, as the spatial dimensions seemed to have no impact on reading times. In a follow-up experiment, participants were asked to memorize a map featuring locations from a text prepared by the researchers. They hypothesized that changes in location should trigger a processing cost and thus an increase in reading times, and this is what they found. In a third experiment, the same text was used but participants were not shown the map previously. Reading times were once again unaffected in this third variation, as in the first. They conclude that readers do not generally seem to keep track of the locations of objects unless they are instructed to do so quite overtly (Zwaan, Radvansky, Hilliard, & Curiel, 1998). These results were supported by other studies, including Zwaan, Langston & Graesser (1995) and Therriault, Rinck & Zwaan (2006).

I decided to focus on the three dimensions I considered primary: causation, space and time. Rather than self-paced reading tasks, I wished to be able to freely re-read words within sections and thus have more access to contextual information within them. For this reason eye-tracking was used. The underlying assumption behind eye-tracking is the same as with the reading task used by Zwaan et al. (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998), that reading times are indicative of processing effort. Unlike simple reading time measures however, eye-tracking can give a fine-grained account of individual reading times for single words and differentiate between the very first reading time and repeated readings of words within a sentence. For this reason, rather than again replicating a well-supported study in the same manner, this experiment was designed in order to study the predictive model dimensions through logical inconsistencies within each dimension. As per the event-indexing model, the less subsequent events overlap, the more cognitive effort should need to be expended. There has been no eye-tracking study that has confirmed this aspect of the model. The hypothesis however aligns quite well with predictive model theory, which would predict that changes in how easily events can be integrated with each other within the predictive models would lead to changes in cognitive effort invested. It has also been entirely untested to what a degree, not simply difficulty in integration but direct logical contradiction and problematic descriptions of events impact this model. Finally, for the purpose of testing how predictive models work within actual fictional literature, I believed it was necessary to use a real text rather than one constructed for the study. Since this made it impossible to select material which would contain convenient test conditions of event integration, a natural text was selected and manipulated with contradictions within

the descriptions of causality, time or space of events to discover if these would lead to reading time changes.

The experiment was designed to use a natural text for this purpose, which added an additional layer of challenge. Many experiments in the literature use texts crafted by the researchers to contain specific information as part of the experimental condition. I fully understand and agree with the benefit of doing so for discovering specific effects and bringing these to the forefront. It was however important to me to find the reading time effects that I believe should come about because of predictive model processing in a natural text. In order to find the text I would ultimately settle on I determined that in order to accomplish an experimental timeframe of less than an hour, allowing for set-up, time to speak to participants and other factors a text with a reading time of between 20 and 30 minutes would be ideal. Thanks to having worked as an assistant in the University of Nottingham eye-tracking lab at the time, performing experiments for other researchers I was intimately familiar with both the equipment and the way participants routinely responded to various experimental conditions and this proved invaluable in planning. I set out to find a short story that would conform to the various attributes I needed. In order of priority these were length, genre and the amount of entirely fictional concepts contained in the narrative. I wanted to avoid texts which were too extreme examples of a particular genre, for example horror or romance, as I feared these would favour readers with more experience with these genres and skew the results towards them. I also needed a text in which it would be possible to hide manipulations, as I already knew that the primary experimental condition would be to add changes within individual sentences of the text which I believed would be difficult for a predictive model to resolve. As a result the text would have to be at once relatively neutral in terms of fictional events and concepts, but removed from reality to a degree where strange sentences would be acceptable at face value and not lead participants to catch on. It proved a daunting task, but after reading through many dozens or short stories and anthologies while timing my reading times of them, I found a story that could fulfil my strict requirements. This was the short story *The Second Bakery Attack* by Haruki Murakami (2003, 2011).

In the following I will present the experiment, the participants I gathered as well as all of the pertinent information about how the text was adapted and the experiment built. I will then present the results in full in section 4.2, before moving onto my discussion of them in sections 4.3 and 4.4.

Participants

Overall 40 participants were included for final analysis. A further five were recorded and excluded, four of them based on poor recording quality and one on a technical fault with the eye-tracking software. Participants were a mixture of University of Nottingham first year English students taking part in experiments for course credit, general University of Nottingham students and staff recruited via internal mailing lists and some general members of the public recruited via the website callforparticipants.com. Participants were selected to be native speakers of English but otherwise not chosen for any particular traits. All participants except for those receiving course credit were paid an inconvenience allowance.

Materials

The material used was an adapted version of the short story *The Second Bakery Attack* by Haruki Murakami, taken from the collection *The Elephant Vanishes* (2003, 2011). It was selected because it was approximately 4500 words long which allowed it to be easily read in a one hour experimental session. It is written clearly and engaging but does not make use of many fictional tropes and it is difficult to define the exact genre of the story. It is about a married couple of unspecified age who wake up in the middle of the night being irrationally hungry. As they consider their options, given that they do not have enough food in the house to be satisfied, they get onto the topic of other times where strange things have happened to them. The husband admits to an episode in his past which he never disclosed to his wife, in which he and a friend living in poverty attempt to rob a bakery. The robbery is unsuccessful, as the baker instead convinces the boys to listen to an album of music from Wagner operas, then gives them all the bread they want. Convinced that this is the cause of the problem, the wife insists that they must now successfully perform an actual robbery to rid themselves of their strange hunger. Unable to find a bakery in the middle of the night, the pair instead rob a McDonald's restaurant, taking thirty Big Mac sandwiches in lieu of bread. After fleeing to consume their stolen food, we learn that they indeed feel a relief of their hunger. Due to the slight surrealist nature of the text, which is quite humorous while using simple language and no unusual locations or conflated cast of characters, I deemed it suitable for subtle manipulations which would not make it too obvious that the text had intentionally been manipulated.

The text was split into 95 sections of text to be presented one screen at a time, with lengths of roughly 70-100 words each. Of those 95 sections 30 were selected for manipulations

within two versions of the experiment, henceforth version 1 and version 2 with a third non-manipulated version of the text serving as a control, henceforth version 3, which would allow for a comparison of reading times. Manipulations were split into six classes and numbered 1 to 5 as there were 5 examples of each manipulation included in the text. They were Causality previous 1-5; Causality following 1-5; Time previous 1-5; Time following 1-5; Space previous 1-5; and Space following 1-5.

Manipulations classed as "previous" contradict something which occurs previous to the manipulated word/s, while 'following' contradicts something which occurs later on in the screen. For instance the manipulation "Causality previous 1" from the first version of the experiment:

> We took turns opening the refrigerator doors and hoping, but no matter how many times we looked inside, the contents never changed. Beer and onions and butter and dressing and deodorizer. It might have been possible to sauté the onions in the butter, but there was no chance those two shrivelled onions could *fill our empty fridge* (original text: fill our empty stomachs). (adapted from Murakami, 2011: p. 37)

This manipulation is classed as "previous" as the manipulation suggests the fridge is empty, which clashes with the information a reader has received prior to this that there are contents inside the fridge, and on a causal level it causes a contradiction with the usual expectation that cooking food (to sauté the onions in the butter) would lead to it being eaten, not put back into an empty fridge.

Manipulations classed as "following" contradict something which follows the manipulation in the text, as for example in "Causality following 2" from the second version of the experiment:

> "That's not true." She looked right at me. "You can tell, if you think about it. And unless you, yourself, personally *intensify* (original text: break) the curse, it'll stick with you like a toothache. It'll torture you till you die. And not just you. Me, too." (adapted from Murakami, 2011: p. 41)

In this manipulation, the character's dialogue is changed to speak of intensifying the curse, which clashes with our usual causal expectations in the following sentence as they go on to describe how the curse will continue to torment them. As the phrasing is "unless you […] intensify the curse" a normal prediction would be that the result should be that the curse should not continue on, or even worsen. Each manipulation is labelled and numbered 1-5 for the purpose of this analysis. Care was taken so that each contradiction was related to

the screen in which it appears as far as possible, meaning that participants did not generally have to memorize past screens in order to notice them – or potentially resolve them. Each participant read only one version, and participants were not made aware other versions existed until they had completed the experiment.

Unfortunately, due to an error during the coding, an early design iteration of version 2 was used for the experiment, which does not contain exactly 5 of each manipulation but a different distribution of 6 causality previous manipulations, 3 causality following, 5 time previous, 4 time following, 7 space previous and 6 space following . After much reflection I decided to include these results as well, as they also proved interesting and gave additional data points to version 1 which together with the control version provided enough data to draw conclusions.

Following the eye-tracking, a pen and paper survey was given to participants. The final, most important, section asked participants to describe whether they felt that there had been anything strange about descriptions of causality, time and space in the story as three separate questions with room for participants to freely answer (for the full survey see Appendix A). Participants were encouraged to answer as fully as they wished and to offer specific examples if they could recall any. A number of other questions were asked to collect data for future analyses of the dataset, which are not included in this thesis.

Procedure

Recording was done using an SR Eye-Link 1000 attached to a recording PC and screen together with a laptop featuring an eye-tracker interface for the researcher. The eye-tracker recorded at 1000Hz, while participants were seated exactly 70cm from the display screen, using a chin rest to minimize head movements. Participants were informed of the purpose of the experiment and the overall procedure and health and safety concerns, then gave informed consent. Participants were instructed that they would be reading a story of undisclosed genre and length and asked to read it as if reading at home for pleasure. This set up took an average of 10 minutes. Each participant read only one version of the text, and no participants were informed until after the experiment that there were any alternate versions of the text they read.

Text was presented one section at a time, averaging approximately 100 words, triple spaced in size 16 font to ensure comfortable readability and more accurate recording. Participants read each section at their own pace and pressed a key on a keyboard in order to move on to the next screen. Questions asking about aspects of the text occurred

between screens and were inserted to ensure readers' attention and that the story was being read for comprehension. All comprehension questions were yes/no questions and participants pressed the n key for no or the y key for yes responses. The eye-tracking part of the study took an average of 30 minutes.

After each participant completed reading the entire text, they were given the pen and paper survey which took an average of 5 more minutes, and were given the opportunity to ask questions and give general feedback as well as receiving their inconvenience allowance.

After exclusions, fourteen participants read version 1, twelve participants read version 2, and fourteen participants read version 3.

Predictions

My predictions for this experiment, given the work of Zwaan et al. (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998) as discussed above, and on predictive model theory as outlined in section 1.5 were to find significant changes in reading times for manipulated sections. The hypothesis was:

If readers use predictive models to process a text and build up full predictive models and expectations based on their global model, then disruptions of expectations and of predictive models should change reading times between the experimental condition and an unchanged control text as readers adapt their predictions.

Based on the work of Zwaan et al. (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998) and other eye-tracking literature, the change in reading times is likely to be an increase in reading times for manipulated sections.

If the null hypothesis is true, then reading times will not change between the experimental condition and the control.

I believed that the majority of contradictory or confusing changes to the wording of each manipulated section should lead to increased error and in turn a modulation of the predictive model formed by a reader, leading to further stages of error suppression, which would be reflected in different fixation times. These should be noticeable at the level of the section of text. Since the manipulations affect the entire meaning and logical coherence of each section, those sections containing manipulations would be read differently. For the same reason, I predicted that it would be unlikely for the actual manipulated words

themselves to necessarily show an effect, as the contradiction requires all of a passage to be read in many cases, particularly for manipulations classed as following, which only show a contradiction within information read after the manipulated words themselves. I considered it a possibility that even the non-manipulated sections within a manipulated version would also show some effects of overall increased difficulty in following the narrative and that this would also lead to changed reading times. This would be tested in the analysis by comparing the manipulated sections from each version against the control in various stages, beginning with all manipulated sections, then moving into more fine-grained analyses of specific groups of manipulated sections. Mean reading times given with standard deviations are per interest area, where interest areas were full, individual words or contractions. The results of this analysis are presented in the next section.

4.2 Results

In this section I will present all of the fully analysed results of the experiment. There will be no formal discussion of the data at this point and the chapter is intended to form a point of reference for the subsequent detailed discussion to follow in Section 4.3 and 4.4 and the rest of the thesis. This section is split into several analyses performed on the data, reported separately in order to avoid confusion.

Before analysis, the data was cleaned using the SR Data-Viewer software's built in 4 stage process, discarding fixation times below the threshold of 80ms and merging some smaller fixations within an interest area. All results were calculated using dwell times as reading times corrected for string length differences between versions. Dwell time was chosen as I wished to account for differences between the total reading times of sections as time spent by a participant before moving on to account for their full reading and comprehension process. This means total dwell times in each interest area, with each interest area being defined as one full word or contraction within each section, were divided by the number of characters contained in the area in order to account for the manipulated texts having slight differences in the amounts of words to the control. This results in a value of reading time per character, which can be compared between datasets even when the word amounts of the samples slightly differ. For the purposes of linear regressions used to test if results were statistically significant, values were then log transformed, allowing for an accurate linear regression calculation.

Overall, there was a significant difference in total reading times between the three versions of the text, confirming the main hypothesis. Surprisingly however, the reading times for the experimental versions were not slower, but faster. Version 1, a manipulated version, had a mean reading speed per character of 62.9ms (SD = 50.9ms). Version received a mean reading speed per character of 65.4ms (SD = 55.2ms). Version 3, the unmanipulated control, received by far the highest value, or slowest reading speed with more time spent per character, at mean 71.4ms (SD = 61.1ms). Note that while I said above fixations of lower than 80ms were discarded, the values reported here fall below 80ms. This is because the original data was recorded measuring fixations on whole words, while the calculation divided these by the number of characters as explained above, in order to account for the difference in length. A linear regression of the log transformed values shows that there was a statistically significant difference in these reading speeds when comparing both Version 1 against the control, and Version 2 against the control, as given in table 1. The linear regressions and all regressions to follow were modelled with the log transformed reading time per character as dependent variable, and version as independent variable.

Table 1: Linear regression of log dwell times per character by version

	Estimate	Std. Error	t value	Pr(>|t|)	
Version 1 vs 3	0.051349	0.003314	15.49	<2e-16	***
Version 2 vs 3	0.080168	0.006795	11.8	<2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Given my prediction this was an unexpected result, but one which was robust across a variety of analyses. Below I shall present a detailed analysis of reading speeds and regressions performed on the various manipulations and subsets of them on the dataset for version 1 as compared to the control, followed by a breakdown of analyses performed on version 2 in comparison to the control which did serve to confirm the results of version 1.

Analysis: Version 1

For the initial detailed analysis, only manipulated sections of text from version 1 were compared to the equivalent, unmanipulated sections of the control. For instance if section 11 was manipulated in version 1, this was compared to section 11 from control version 3. The same method of comparison was used throughout the analysis. The aim was to discover in more detail if all manipulated sections overall led to a change in recorded reading times. Across all manipulated sections, version 1 reading times per character were M = 65.2ms, SD = 52.9ms. Reading times per character for the equivalent sections of the control were M = 71.9ms, SD = 63.4ms. Here too, the control was read more slowly, with more time spent per character by participants. Linear regressions of the log transformed data confirmed this to also be a statistically significant difference as reported in table 2 below.

Table 2: Linear regression of log reading times per character by version

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.876817	0.013258	292.41	< 2e-16	***
Manipulations vs V3	0.037580	0.005836	6.44	1.22e-10	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Next, the manipulations were investigated more thoroughly by type. They were broken down and analysed in terms of both the following and previous type, as well as both combined for each predictive model dimension. In each case only the sections of text containing the manipulations and their equivalent sections from the control were compared. This same method was used for the remainder of manipulation analyses following.

Beginning with causality, all causality following types were grouped, as well as all causality previous types, and finally both combined to represent all causality manipulations in version 1. These were once again compared to their equivalent sections in version 3, the control. Reading times are reported below in table 3.

Table 3: Reading times per character for causation manipulations vs control

	Mean	Standard Deviation
Causality following	67.0ms	53.6ms
Control	69.5ms	59.8ms
Causality previous	64.8ms	55.3ms
Control	71.9ms	60.3ms
Causality total	65.8ms	54.6ms
Control	70.9ms	60.1ms

Linear regressions showed that interestingly, the causality following condition, manipulations which were intended to clash with information following the manipulated word or words, did not reach statistical significance in terms of reading difference. Causality previous conditions did, and all causality manipulations compared to the control overall also did reach a statistically significant difference. All values are reported in table 4.

Table 4: Linear regression of log reading times per character for causality manipulations per type by version

	Estimate	Std. Error	t value	Pr(>|t|)	
Casuality following	0.02145	0.01557	1.377	0.168	
Causality previous	0.04054	0.01231	3.293	0.001	**
Causality total	0.032900	0.009679	3.399	0.00068	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Given the failure of causality following type manipulations to elicit an effect, these were investigated more throroughly, with linear regressions run for each individual manipulation using data from only the relevant section of text and control. These results are reported in table 5. These show that of the five different manipulations, only causality following 5 led to a statistically significant reading time change, with manipulated text for the only time in this study being read more slowly (M = 76.1ms, SD = 66.4ms) than the control (M = 62.4ms, SD = 54.2ms).

Table 5: Linear regression of log reading times per character for individual causality manipulations per type by version

	Estimate	Std. Error	t value	Pr(>|t|)	
Casuality following 1	0.04087	0.03390	1.206	0.229	
Casuality following 2	0.05930	0.03433	1.727	0.0846	
Casuality following 3	0.08645	0.05336	1.62	0.106	
Casuality following 4	0.02132	0.02627	0.811	0.417	
Casuality following 5	-0.07157	0.02558	-2.798	0.0053	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Next, space type manipulations were analysed in the same way, also showing a clear difference in reading speeds in which the control had more time spent per character by participants. Reading times by manipulation and equivalent control sections are reported below in table 6.

Table 6: Reading times per character for space manipulations vs control

	Mean	Standard Deviation
Space following	60.7ms	50.5ms
Control	67.8ms	61.0ms
Space previous	66.7ms	55.8ms
Control	72.8ms	67.0ms
Space total	63.6ms	53.2ms
Control	70.2ms	64.0ms

Linear regressions of the log transformed reading timer per character showed a statistically significant difference for all three comparisons, as reported in table 7.

Table 7: Linear regression of log reading times per character for space manipulations per type by version

	Estimate	Std. Error	t value	Pr(>|t|)	
Space following	0.03946	0.01501	2.628	0.00862	**
Space previous	0.02943	0.01334	2.207	0.0274	*
Space total	0.03464	0.01011	3.427	0.000614	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Finally, time manipulations were analysed by type and compared to the equivalent sections in the control. Once again, reading times per character were higher in the control sections. All values are recorded below in table 8.

Table 8: Reading times per character for time manipulations vs control

	Mean	Standard Deviation
Time following	66.9ms	51.0ms
Control	72.7ms	61.5ms
Time previous	65.4ms	50.4ms
Control	76.7ms	70.4ms
Time total	66.2ms	50.7ms
Control	74.7ms	66.1ms

Linear regressions of the log reading times per character against the equivalent control sections indicated that all three comparisons were statistically significant, as reported in table 9.

Table 9: Linear regression of log reading times per character for time manipulations per type by version

	Estimate	Std. Error	t value	Pr(>|t|)	
Time following	0.02990	0.01243	2.404	0.0163	*
Time previous	0.06262	0.01704	3.675	0.000242	***
Time total	0.04622	0.01056	4.378	1.22e-05	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Overall the results for version 1 when compared to our unchanged control text showed a robust effect of the manipulations on reading times, with longer reading times per character in the control under all conditions but one. Under individual analysis only causation following type manipulations, changes which affected portions of a text following the manipulation itself, failed to reach statistical significance in terms of reading speed differences as a group, with a single causation following type manipulation being read statistically significantly slower. Next I will present the analysis of the data gathered from version 2 of the experiment which was also manipulated.

Analysis: Version 2

.As reported at the beginning of the results section, overall reading time were faster for version 2 than the control, and this was statistically significant. For the further analysis, all manipulated sections of version 2 were compared to their corresponding sections from the control version 3 to see whether there was a significant difference in reading speeds as in version 1. Data was transformed in the same way as these two versions also differed slightly in numbers of words and letters in some section, so once again reading time per character was chosen for a truly fair comparison, and log reading time per character for linear regressions. Reading times for manipulated sections of version 2 (M = 65.1ms, SD = 55.7ms) were also faster than for the corresponding sections from the control (M = 69.0ms, SD = 58.0ms). A linear regression of log reading time per character showed that this was statistically significant as in table 10 below. As in the analysis of version 1, all linear regressions were run using log transformed reading times per character as dependent variable and version as independent variable.

Table 10: Linear regression of log reading times per character for manipulated sections vs equivalent sections in the control by version

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.75165	0.03213	116.757	< 2e-16	***
Manipulations vs v3	0.06263	0.01247	5.021	5.18e-07	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Next, each type of manipulation was analysed as in version 1, by type of manipulation as well as predictive model dimension. The dataset was split up so that each type of manipulation analysed used a dataset of only the manipulated sections, and only their equivalent sections in the control. I shall begin with causation manipulations, for which reading times per character were faster in the manipulated version 2 sections than in the control, reaffirming the general results of version 1. Values are reported in table 11.

Table 11: Reading times per character for causation manipulations vs control

	Mean	Standard Deviation
Causality following	65.6ms	71.4ms
Control	68.5ms	59.8ms
Causality previous	61.7ms	46.3ms
Control	64.3ms	52.0ms
Causality total	63.5ms	58.9ms
Control	66.2ms	55.7ms

Linear regressions showed that for this version of the experiment, the log reading time per character differences between manipulated text and control failed to reach significance for the causality previous type manipulations, which were designed to clash with information obtained just before the manipulation, and further analysis showed this to be the case for all such manipulations individually. Differences did reach statistical significance in the causality following conditions, and also when taking all causality manipulations into account for a comparison with the equivalent sections in the control as in table 12.

Table 12: Linear regression of log reading times per character for causality manipulations per type by version

	Estimate	Std. Error	t value	Pr(>|t|)	
Casuality following	0.09237	0.04121	2.241	0.0251	*
Causality previous	0.02575	0.02607	0.988	0.323	
Causality total	0.05571	0.02344	2.376	0.0175	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Next, space manipulations were analysed using the same method as previous manipulations, with reading times also showing lower values for manipulated sections than in the equivalent control sections as reported in table 13.

Table 13: Reading times per character for space manipulations vs control

	Mean	Standard Deviation
Space following	61.5ms	48.4ms
Control	66.5ms	50.9ms
Space previous	69.0ms	57.0ms
Control	71.7ms	58.9ms
Space total	65.6ms	53.3ms
Control	69.3ms	55.4ms

A linear regression of log reading time per character showed an interestingly similar result to causation manipulations, with following type manipulations and space manipulations in general reaching statistical significance, but previous type manipulations in isolation failing to differ significantly from their equivalent sections in the control as reported in table 14.

Table 14: Linear regression of log reading times per character for space manipulations per type by version

	Estimate	Std. Error	t value	Pr(>|t|)	
Space following	0.09075	0.02783	3.261	0.00112	**
Space previous	0.03734	0.02513	1.486	0.137	
Space total	0.06217	0.01869	3.327	0.000881	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

An individual analysis showed that individually, two out of the five space previous manipulations did reach statistical significance. Interestingly, of these two, the first, space previous 1, was read faster (M = 59.3ms, SD = 45.4ms) than the control (M = 69.9ms, SD = 58.3ms). The other statistically significant condition, space previous 2, was in fact read more slowly (M = 79.0ms, SD = 65.8ms) than the control (M = 59.0ms, SD = 41.7ms). Linear regressions for the values are reported in table 15.

Table 15: Linear regression of log reading times per character for space manipulations per type by version

	Estimate	Std. Error	t value	Pr(>|t|)	
Space previous 1	0.1784	0.0795	2.245	0.0251	*
Space previous 2	-0.21339	0.08568	-2.491	0.0135	*
Space previous 3	0.02546	0.05892	0.432	0.666	
Space previous 4	0.01048	0.06186	0.169	0.865	
Space previous 5	0.04412	0.06576	0.671	0.503	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Finally, time manipulations were analysed by comparing reading times per character, also displaying the trend of faster reading speeds in manipulated sections compared to the control, as reported in table 16.

Table 16: Reading times per character for time manipulations vs control

	Mean	Standard Deviation
Time following	70.7ms	58.8ms
Control	70.9ms	60.7ms
Time previous	61.4ms	50.2ms
Control	72.3ms	65.7ms
Time total	66.2ms	55.0ms
Control	71.6ms	63.2ms

Linear regressions of the log reading times per character against the equivalent control sections indicated that time following type manipulations failed to reach a statistically significant difference in reading times per character to the control, and individual analysis of time following manipulations showed that none of the individual manipulations presented statistically significant differences from the control. Differences between time previous type manipulations and time manipulations in total were statistically significant, as reported in table 17.

Table 17: Linear regression of log reading times per character for time manipulations per type by version

	Estimate	Std. Error	t value	Pr(>|t|)	
Time following	0.006563	0.035414	0.185	0.853	
Time previous	0.13629	0.02985	4.566	5.17e-06	***
Time total	0.06932	0.02330	2.976	0.00294	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

These results are in agreement with the results of version 1. The manipulated sections of version 2, just as the entire version 2 text, were read more quickly than the equivalent sections and entire text of the unchanged control, version 3. All three types of manipulations taken in their entirety presented statistically significant differences in reading speed per character between manipulated text and control. Unlike in version 1 however, more types of specific manipulations in isolation failed to reach statistical significance. While this affected only causation following types in version 1, for version 2 causality previous, space previous, and time following manipulations themselves did not present statistically significant reading speed differences from the control. I initially believed that this was possibly due to the fact that rather than the intended balance of 5 of each type of manipulation, version 2 accidentally included a different balance, but this would in fact appear unlikely. Due to my mistake, version 2 contained 6 causality previous manipulations, 7 space previous and 4 time following manipulations. More detailed analysis showed that only two individual conditions, both of the space previous type, reached statistical significance in terms of reading speed differences, with one interestingly being faster than the control as per general manipulations, and one being slower. Only time following, which did not achieve statistical significance, was one experimental trial short, while the other two which did not reach statistical significance in this version of the experiment were in fact overrepresented.

In summary, based on the overall analysis and the outcomes of all manipulation types combined across both experiments, I concluded that the main hypothesis was supported, and the manipulations made to the text caused statistically significant reading time differences against a control text. In a great surprise, the secondary hypothesis, that manipulated text should be read more slowly, was falsified. The result in all but two individual cases was that the control was read more slowly than manipulated text.

I shall now present the analysis results of the post eye-tracking survey before moving into a discussion of the results.

Survey Analysis

The post-eye-tracking survey asked several questions whose participants responses to which I will not report in this section, as they were mainly intended as additional data points if necessary, or for future alternate analyses of the data gathered (the survey is available entirely in Appendix A). While some of the information provided by participants on their general time spent reading various materials was tested as predictors for the main analysis, there were no significant interactions. The most interesting and relevant data which was collected primarily for this analysis was collected in the final section of the survey, in which participants were asked to provide free form answers as to whether they believed there had been anything strange in the text in regards to descriptions of time, space and location, and cause and effect relationships.

The responses were individually encoded as 1-3. One corresponds to responses which indicated that a participant thought there was nothing wrong with the respective dimension of the story. Two responses correspond to responses in which participants reported feeling that aspects of the story were strange, but with no reference to manipulated sections or any indications of awareness that there were manipulations. Three represents responses in which participants explicitly pointed out manipulations. These results are reported below in Table 18.

Table 18: Survey responses for causality, time and space by version and code type

Version 1

Code type	time	space	causality
Code 1	6	4	4
Code 2	6	9	10
Code 3	2	1	0

Version 2

Code type	time	space	causality
Code 1	6	4	4
Code 2	4	6	7
Code 3	2	2	1

Version 3

Code type	time	space	causality
Code 1	4	2	3
Code 2	10	12	11

Overall, out of the 120 responses (40 participants x 3 choices), the vast majority were a code 2, indicating the participants overwhelmingly considered the story to be strange on all three aspects of causality, time and space but that this was not based on the manipulations, at least on any conscious level. 75 out of 120 or 62.5% felt this way. Regarding the reports of manipulated sections, we must take into account that Version 3 did not contain manipulations and this could not have led to any code 3 responses. 26 participants in total read Version 1 and 2, for a total of 78 possible responses, out of which eight were a code 3, representing 10.7%. Time manipulations appear to have had the most conscious impact, receiving two code 3 responses for both versions 1 and 2, while causality manipulations went entirely unnoticed in version 1, and received a single report in version 2. Based on this I concluded that the experimental materials worked as intended, without manipulations being overtly obvious and pulling the validity of the experiment into question, and showing that participants in the control appeared to overwhelmingly consider the text unusual with no manipulations. I will return to choice free form comments which were of particular interest in the discussion, which will be the next section. I will discuss these intriguing results and offer an argument based on the fact that this appears to be a novel result in eye-tracking, in which an increased difficulty in a text lead to a speeding up rather than slowing down effect in terms of reading times.

4.3 Discussion

In this section I will discuss the results of the experiment, their relevance for predictive model theory, and then begin to offer an explanation for the observed effect. I shall return to some earlier textual examples from Section 3.4 to relate these to the text of the experiment and show what I believe may have happened for readers of my experimental text.

The overwhelming result of all four analyses is that there were statistically significant changes in reading times between the experimental conditions and a control, with the manipulated text in both conditions having been read faster than the control. This was the case for both the manipulated sections of text themselves, as well as the texts in their entirety, indicating a quite robust effect. While this confirmed one part of my prediction, it was contrary to the entailment of my prediction, as I had hypothesised that reading times should have increased for manipulated sections, not the control. The faster reading speed relative to the control was however generally robust across all three classes of manipulations which corresponded to the predictive model dimensions. The fascinating exception was the causation following condition of version 1, as well as space previous 2 of version 2 which remarkably were the only conditions to not show a faster reading speed, and which appeared to in fact elicit the hypothesized reading speed decrease. Given that this was a case of one against twenty-five other manipulations which showed a statistically significant reading speed increase for version 1, and one case against fifteen which showed a statistically significant speed increase in version 2, these had to be considered as outliers upon which it is difficult to draw conclusions.

This was a very surprising result. After much consideration and reflection, it was clear that this did not constitute the null hypothesis being true as there was a definitive difference in reading speed which could not be explained by participants noticing the experimental manipulations or other factors, and made me reflect on why I had necessarily expected reading times to increase in all cases, and why I accepted the hypothesis of higher difficulty unilaterally equating longer reading times as a given. Following this reflection, I do believe that these results support the principle of predictive model theory as applied to the reading process, and in fact lend more support to the importance of global model integration than I had at the time understood, as the manipulations did have an overall effect on the reading times in both manipulated versions. A key conclusion to be made from this is that as I predicted within the one-to-one principle in section 1.5, the predictive model dimensions

are equally important in influencing overall reading speed. Similarly, at which point conflicting information is introduced appears to make little difference when conflicts are spread relatively equally throughout a narrative. Coupled with the clear lack of responses indicating much conscious awareness of the manipulations, there was clearly a difference in the way participants processed the manipulated text and the control. I believe that there were indeed errors elicited by the manipulations, but that these errors were resolved within either the predictive model stage, or within the global model integration but before the full representation was formed. Participants overwhelmingly pointed out that they considered the story to be strange, likely activating certain global model expectations about the particular kind of fictionality portrayed, and in turn chose to normalize many of the manipulations. Essentially, once the global model level of interpretation is that the narrative is strange, confusing or surreal, then contradictory information elicits errors which are far more easily explained by the existing attitude that this is expected. That is, since the actual content of the global model and the predictive models integrated into it as the conceptual priors contain the outlook that the story is nonsensical and difficult to understand, contradictory and difficult to understand sections generate less serious errors and integrate into this expectation more easily.

Once participants have fully come to the conclusion that the text is surreal and illogical, the top-down interpretation of the global model and full predictive models would lead to an overall attitude that the text is predicted to make little to no sense. Once such representations are formed around the initial manipulations, they will modulate what counts as an error in the first place. Using such predictive models as priors, the manipulations would not elicit any new errors at all, but simply fulfil expectations of implausibility.

Given the constancy of the difference in reading times between manipulated text and control, coupled with participant responses, I believe that this explanation is able to account for the differences. This is further supported by the results of the survey questions given to participants regarding the text. One clear thread running through the responses was confusion regarding a boat metaphor found within the narrative. In it, the protagonist compares the irrational hunger he feels to being in a boat floating on an ocean above an underwater volcano which is erupting and gradually growing closer. The story closes with it being gone, and him finding himself floating on calm waters instead, signifying the end of the hunger. Despite being an integral part of the story and the way it is told, many participants appeared to struggle with it. Some choice comments include:

- “The description of the lake was strange as it was not a literal place, just a place in the protagonists brain referred to by his curse.”
- “I think the boat metaphor was a bit unnecessary, and there should have been more description of the setting […] their hunger seems unrealistic on the basis that they are clearly within distance of many shops and take-aways […].”
- “The description of the imaginary location in the protagonist’s head was strange and dreamlike.”
- “The boat metaphor was strange.”
- “[…] the mention of the boat, and certain bodyparts e.g. the stomach seemed so real you had to remember they aren’t.”

The desire of participants to interpret the story as close to reality as possible is further strengthened by considering the responses which pertain to causality within the story, such as:

- “[…] it wasn’t explained why there was a curse and why those specific actions needed to be taken in order to lift it, or why they were so hungry.”
- “I couldn’t understand the correlation between the bakery and McDonalds.”
- “[…] how was the first attack on the bakery being a curse could be explained to give a better understanding.”
- “Logic behind robbing the bakery did not make sense to me.”
- “Difficult to determine fiction from reality – what relevance does the bakery have?”

These comments further betray that these five participants were attempting to discover a fairly realistic cause and effect relationship within the story, but were unable to do so. As a result, it is possible that these participants became unable to differentiate between manipulations and other aspects of the story which they also felt to be implausible, thus no longer eliciting errors. They were not aware of this being a story by Murakami or the way he generally writes surrealistic fiction anchored in familiar and otherwise realistic settings which amplify the absurd humour embedded within the narrative. This may have led participants to use the fairly normal and descriptive opening of the story as justification for reading it as a realistic story, further strengthened by the use of real locations around Tokyo and the quite famous and real fast food restaurant McDonald’s. What the participants did in response to this was to adopt a global stance that the narrative could not be followed, and accepted the absurd elements, amplified by my manipulations, at

face value. They were explained by fitting into an illogical narrative and did not require further processing.

Meanwhile, in the control version, the story was still absurd but the text was more internally coherent. Each screen followed a mostly logical sequence, and importantly, this required more effort on the part of the participants because they likely began to store these logical sequences in their working memory. The manipulation “space 5 previous” perfectly showcases this. In the original story, the protagonists are driving away from the restaurant, for about half an hour. They are then described eating some of the stolen hamburgers, and we are told what numbers of hamburgers remain behind on the back seat. For the participants who were given the manipulated version, the hamburgers were left behind at McDonald’s. Adopting a global strategy that the story is difficult to follow, they would likely not have questioned this much, and either accepted that some hamburgers were left behind for an unknown reason, just as many other things happened for unknown reasons due to the manipulations, or accepted the close semantic link between hamburgers and McDonald’s.

For the participants reading the control, they had to consider several additional things. Firstly, they were not previously told that the hamburgers were left on the back seat, or put there. They had to infer that they were put on the back seat before or after the protagonists ate some. Then they had to potentially think back to previous screens to work out if the number of hamburgers left on the back seat tallied with the number eaten and the total number stolen in the first place, requiring activation of working memory. In essence, I believe something similar to this happened in all cases. It is not the case that nobody noticed the manipulations at all. Clearly participants were aware of the increased logical consistency and event overlap and were directly affected in terms of their reading behaviour.

Manipulations which did not follow this pattern remain of interest as well. Manipulations such as the causality following types in version 1 and causality previous in version 2 did not result in the same reading speed increase. Instead, these types were read at similar speed to the control, indicating that these manipulations caused as much cognitive effort as the control did for these sections of text. It is possible that the strategy of expecting the text not to make sense is not so rigid that participants could not notice further clashes, and that perhaps some of these manipulations were too egregious, although they did not receive particular attention in comments by participants. Further studies into this area and causal

relationships in particular would be of great interest here. It is already well documented that extreme causal clashes will cause difficulty in existing studies such as those of Zwaan et. al. (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998) and also in studies of mismatch negativity as discussed in section 2.3, but I believe my study indicated an opportunity for discovering the point at which readers may accept problematic or even impossible causal relationships if presented in the context of a narrative and when these become truly unacceptable in any context.

Overwhelmingly, my manipulations in this study proved is that it is not a given that an increased difficulty or textual clash increases cognitive effort and reading speed. The strategies used by readers when expending more or less cognitive effort are decided online, can vary, and readers can chose to expend less effort without "giving up" on reading the narrative, and still consciously discuss passages they found troubling. This is an exciting new discovery: they proved that context is far more important than previously acknowledged in literature on processing and situation model theories, and that in fact it is possible for errors to lead to strategies of predictive model and global model integration which reduce processing costs by modulating what counts as an error against the current predictive model and priors rather than simply generating infinite layers of explanation, instead directly approaching texts with the prediction of reduced plausibility being acceptable, while still remaining sensitive to errors which cannot be integrated. This is to my knowledge a novel discovery in the field. I believe this is a very exciting field for further study, and can envision a repeat of this experiment to find out more conclusively where exactly the threshold might lie at which readers decide to modulate their expectations to reduce future error, or to engage at the cost of further cognitive effort to form further predictive models and actively suppress errors.

In this discussion I have concluded that my hypothesis was supported by the experiment, and that predictive models are used by readers to modulate their expectations, and that this directly affects reading times of texts. I have argued that it is possible for this to lead to strategies of reading more quickly, and modulating expectations in order to reduce overall cognitive effort when it is not critical that we understand everything that is happening in detail while at other times recognizing and processing more serious errors in the same text. This was demonstrated by close analysis of the textual features which may have led to this, and responses by participants in relation to these features.

This directly brings into focus questions of plausibility and the way that we treat questions of what is believable and what is not and how this is modulated by context. In the next section, I shall discuss the notions of linguistic and contextual plausibility which I developed as a direct consequence of the experimental results.

4.4 Contextual Plausibility

Building upon the results of my first experiment, I will introduce the notion of contextual plausibility and linguistic plausibility in this section. I shall first introduce the concepts then discuss their relevance to predictive model theory and how they relate to the experimental results and my discussion in Section 4.3. I will then go on to analyse the short story used for the experiment using the concepts of plausibility I introduce in this chapter as well as a general predictive model analysis to show how it may have affected readers.

What I shall call linguistic plausibility is the perceived "fit" or likelihood that we will believe a sentence based on what our processing tells us that the sentence means semantically and how well it integrates into our global model of the world. Linguistic plausibility was at play in examples of pragmatic normalisation across chapter 2, as well as in some of the examples discussed in section 3.4, where questions of gravity and thermodynamics suddenly interfered in a fictional setting. Without recourse to additional information, comparing those issues to our knowledge of reality is the default. Sometimes, there is more to it than that. Those examples I discussed worked this way in isolation and because of the location in the respective stories where they were from, both appearing fairly early on in the fiction. Let us return to this specific example:

> Annoyed, Richard took the sword in both hands, feeling the anger surge through him. He gave a mighty swing at the remaining tree. The tip of the blade whistled as it sliced through the air. Just before the blade hit the tree, it simply stopped, as if the very air about it had become too thick to allow it to pass.
> Richard stepped back in surprise. He looked at the sword, and then tried again. Same thing. The tree was untouched. He glared over at Zedd, who stood with his arms folded and a smirk on his face.
> Richard slid the sword back into its scabbard. "All right, what's going on."
> (Goodkind 1995, p. 126)

At this point the reader is forced to consider the issue at face value, as I analysed it in section 3.4, and assume that normal Newtonian motion is disrupted, which makes the actions of the sword implausible and introduces error. As readers, we can integrate this

into the next predictive model, in which the character wielding the sword is also annoyed, leading to a predictive model which would handle the error by deducing that the reason for the error is something unknown in the story, where it is also an error. It is implausible that the sword would simply stop mid-swing without anything offering resistance. However, following this, the other character, Zedd, explains that this has to do with the sword being a special kind of sword which has magical properties. He creates a contextual plausibility for us, by explaining that the magic of the sword makes it impossible for the sword to hit or injure anything or anyone whom its wielder does not believe is deserving of death. As a result, in all future instances of this sword being swung, as readers we are no longer relying on integrating instances of the sword stopping with our global model knowledge of motion. Instead, the novel has given us contextual plausibility. If we believe that Richard believes another character deserves to die, it is plausible for it to behave like a normal sword and to harm them. If we however believe that Richard does not think someone is deserving of harm, such as a tree, then it becomes more plausible that the sword stops in mid-swing and spares them. I think that this is a wonderful example because it is a case in which the contextual plausibility actually gives several options for the same turn of events to be plausible or implausible in the same story. It replaces our usual criteria for considering a sentence semantically implausible, i.e. the sword stops suddenly, and offers a new criterion which overrides it making the sentence plausible at times of the author's choosing. Importantly, it only does so within the context of that particular fiction. This is what I call contextual plausibility.

I would furthermore like to suggest that there are different kinds of contextual plausibility, and that these are what make fictional reading and our predictions of how fictional reading processes work difficult. The two types of contextual plausibility which I would like to focus on are *genre based contextual plausibility*, or henceforth *genre plausibility*, and *text-specific contextual plausibility* or henceforth *textual plausibility*. I define genre plausibility as consisting of specific criteria of contextual plausibility which make statements that would normally clash with our beliefs about reality plausible, and which occur across a body of different texts over time, in a variety of media. I call it genre plausibility because I believe that in fact one important aspect of genre is the shared nature of certain contextual plausibility criteria which occur and identify a text as belonging to this family of texts. Genre plausibility will be discussed in more detail in chapter 5.

Textual plausibility I define as a form of contextual plausibility which occurs in a specific text which makes statements that would normally clash with our beliefs about reality plausible, and is not otherwise found in an appreciable body of texts. Both of these definitions are fuzzy, for a variety of reasons. I am quite certain that genres are far more complex than this set of shared plausibility criteria, and also culturally determined, as would the contextual plausibility tied to them be. It is also possible that a text could be considered a part of a given genre even if it shared few, or none of the conventional criteria of that genre's conventional genre plausibility. Textual plausibility meanwhile may in fact be shared across a body of texts by the same author, writing about a persistent fictional universe or simply because it is a stylistic quirk which is successful for this author. Both genre plausibility and textual plausibility may not be in effect for the entirety of every single text, but only account for some sections within in them, or possibly even only one single section in which the normal rules of the world as we perceive it are lifted for a particular purpose, then reinstated. Contextual plausibility may also be tied to an unreliable narrator and lead to further confusion. In many cases, the two may overlap in the sense that certain aspects of a given work's textual plausibility criteria may be shared with genre plausibility. For all of these reasons, I must stress that the two definitions as I have given them above lead to the two categories I discuss as necessarily fuzzy categories. For an individual reader, there will always be a level of knowledge which is entrenched enough to be described as genre plausibility and which the reader is able to use for the purposes of predictions, while other knowledge may be new to them and form textual plausibility. While there is thus no possibility of categorizing any specific textual information as belonging to either type of plausibility in all cases, each individual reader will have a repertoire consisting of genre plausibility against which new textual plausibility may be defined, for that individual, and it is this unique distinction which is helpful to keep in mind when analysing texts using my terminology.

The interaction between these two types of contextual plausibility can help explain phenomena such as the interesting reading patterns which occurred in my experiment, when participants were not made aware of the specific genre of the text they were reading. By necessity, they were forced to weigh whether certain accounts of plausibility in the text with which they were presented were textual and unique to this work, or if they were based upon the genre of the text and of wider criteria of genre plausibility. One of the potentially big tells for a reader is the circumstance as discussed in the example of Goodkind (1995) above: textual plausibility is often made explicit within a text because it

differs from expected circumstances and general genre plausibility. In the above case, another character explains something to the main character which is equally intended for the reader, something called exposition. Exposition is not necessarily about textual plausibility, as it can also simply help anchor a narrative within a certain genre or setting. Some exposition is simply assertive in that sense, stating for example “This is a world in which there is magic”. These statements help to anchor a text within a specific genre and also to locally give readers a good introduction to the kind of story which follows. The important thing about this kind of exposition is that it is not causal. There is no information of why the world contains magic, how it works, how or why it differs from the real world etc.

Exposition of the kind used by Goodkind (1995) for his Sword of Truth novel goes this important step further, and offers an explanation for a phenomenon. It tells the reader “This is contextually plausible in this world *because* of x”. The x here contains elements of magic and some related tropes, which a reader was made aware of earlier in the story. The exposition makes clear that something different is being added to it however, which is the spell that makes the sword act the way it does. It offers a causal relationship between the fantastical elements of the world and the realistic elements such as swords which also exist in the real world. A reader is meant to suspend the normal rules of these objects in favour of the new causal relationship offered by the text, and under the general assumption of readers’ applying the principal of minimal departure (see Section 3.3) this will usually be the case. The major difference between this novel and the short story by Murakami (2003, 2011) which I used in the experiment is that Murakami does not use the same kind of exposition much within his texts and creates striking and often humorous clashes when expectations based on real world knowledge are suddenly and decisively upended.

The Murakami story begins with a seemingly mundane conflict, as both the protagonist first-person narrator who is never named and his wife wake up feeling intensely hungry. The story continues in a dry and factual manner until coming to the metaphor in which the hunger is compared to the narrator sitting in a boat on a lake. This is something which seems to have clashed quite strongly with the general feelings about plausibility most of the participants of Experiment 1 had within the story as evidenced by some of the comments analysed in section 4.3. It likely stems from the fact that Murakami does not offer any account of what kind of world it is or of any new causal relationships. Following Ryan’s law of minimal departure (1995), participants appear to have uniformly assumed

that the world of *The Second Bakery Attack* is our real world. The boat metaphor is the first in a series of fantastical or absurd features which Murakami fits into the narrative without any indication that anything unrealistic or odd is to follow and without offering any account of plausibility or causal relationships. To a reader (let us assume a reader who is generally familiar with literature but has not read Murakami before as represented by the majority of participants) these elements will of course stand out all the more for their suddenness. The protagonist casually admits to having committed robberies with a friend as a young man, rather than getting a job. His wife reacts in an unexpected way, inquiring about the inefficiency of robbing rather than working, without any emotional or ethical concerns. It is then equally odd to learn that the protagonist's wife feels they are cursed. This suddenly introduces an occult element to the story, one which could be taken as a joke, but to which the protagonist does not react in an amused way – he takes her seriously. When they discuss how he listened to an album of Wagner with a baker during an attempted robbery instead of genuinely threatening the man with a knife, the protagonist and the narrative present it as serious. Interestingly, she offers the first new causal relationship in this part, just after suggesting that it was the failed robbery which put a curse on him, and that he should have noticed:

> "That's not true." She looked right at me. "You can tell, if you think about it. And unless you, yourself, personally break the curse, it'll stick with you like a toothache. It'll torture you till you die. And not just you. Me, too."
> "You?"
> "Well, I'm your best friend now, aren't I? Why do you think we're both so hungry? I never, ever, once in my life felt a hunger like this until I married you. Don't you think it's abnormal? Your curse is working on me, too."
> (Muarakami, 2011: pp. 41-42)

Here there is for the first time a concrete discussion of this fictional curse introduced into the story, not in a direct piece of exposition, but through direct dialogue between the characters. The tone of this exchange is not one in which the wife is making an argument. She is stating a list of assertions about this curse and how it will kill both of them unless broken. There is no resistance or doubt in this, which again clashes with the otherwise realistic setting of the story. In the manipulated version for the experiment, the word break in "personally break the curse" was changed to "personally intensify the curse." Upon reflection of this after analysing the results of the experiment and discovering that there was no significant difference in reading speeds between manipulation and original, I believe this is a case of the original text as well as the manipulated text essentially offering a textual plausibility. There is no applicable real world knowledge for a reader to draw on to verify or

deny this statement, and as it fits into the generally surrealist narrative, both versions were essentially taken on board by readers as offered by the text. That is to say, neither version was innately more or less plausible than the other at this stage. Over time however, this textual plausibility begins to fit better into further information given by the text when a reader has been told the protagonist needs to break the curse than when he is told to intensify it. The protagonist seems confused at first, but he goes on to acknowledge this series of statements from his wife without any dissent, only going on to ask her how to go about breaking it. This too, she answers succinctly: "'Attack another bakery. Right away. Now. It's the only way.' 'Now?' 'Yes. Now. While you're still hungry. You have to finish what you left unfinished.'" (Murakami 2011: p. 43). This entire passage rather suddenly and decidedly clashes with a host of real world expectations which readers are likely to have held on to during reading. The realistic setting gives way to one in which curses are real, but which otherwise seems to follow reality. The previously likeable character of the protagonist admits to having committed crimes in his past, which appears to be of no concern to his wife. Her only worry is that he caused them both to be cursed. Without any previous indication, she appears to be an expert on curses and states without hesitation that they need to perform a robbery on a bakery.

In future mentions of the curse, the textual plausibility is that it is real and was caused by the husband's inability to have performed a violent robbery properly. There is a sense of urgency as the curse will kill both him and his wife unless it is broken. Finally, it is clear after this point where the story is going, as it has been established the curse will be broken by successfully performing a robbery. This is aided by some of the language used. All of the causal statements made by the wife are clear and straight-forward assertions. This pattern of causal relationships can be integrated in predictive models surrounding these sections. Mentions of curses and of robbing bakeries as a solution for anything within realistic fiction would elicit errors, and here too the textual plausibility of how these are presented clash with such genre plausibility and a reader is forced to somehow reconcile this during reading as represented in Figure 19.

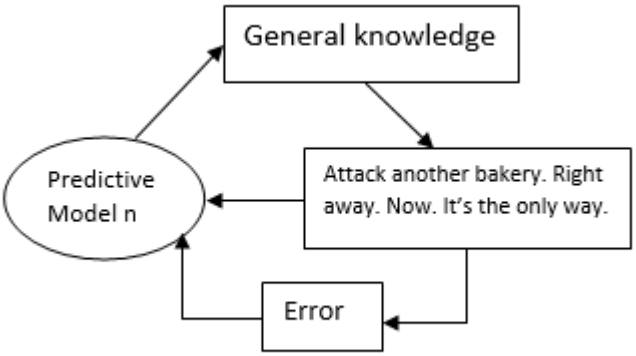

Figure 19: Error elicited from the mention of a curse

With the textual plausibility established by the previous predictive model in which the wife has explained the curse and how to get rid of it, any further mentions of it no longer lead to errors. This textual plausibility is further strengthened when the wife later insists that they must target a McDonald's because: "'It's *like* a bakery,' she said. 'Sometimes you have to compromise. Let's go'" (Murakami 2011: p. 44). The added element which can be inferred from this addition is that the curse may also be lifted by attacking a McDonald's, as apparently compromises are okay and the McDonald's is similar enough to the concept of a bakery. This additional factor becomes relevant towards the end of the story when the pair is questioned by the McDonald's staff about why they have to rob them. The wife offers: "We're sorry, really. But there weren't any bakeries open. If there had been, we would have attacked a bakery" (Murakami 2011: p. 48). Given the prior knowledge established above, this would not lead to any error, even though it makes no sense compared to our real world knowledge, because the priors can account for this statement as in Figure 20.

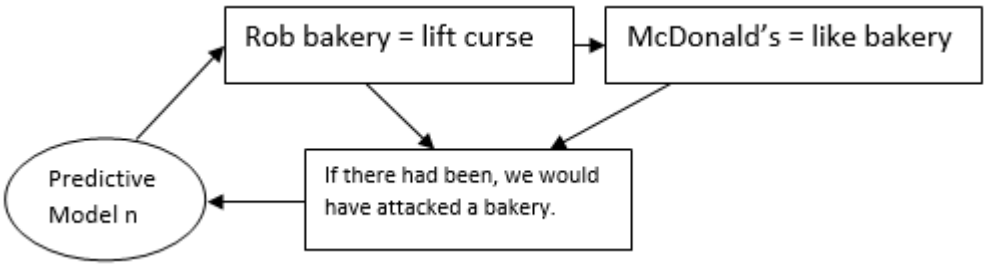

Figure 20: By utilizing prior textual plausibility, no error is elicited

While this use of recent predictive models as priors can lead to the text being easily integrated into the new predictive model, it is not so easily integrated into the global model. The general background of a realistic story is still there, only with additional

elements, but it is not clear which elements. The addition of the robbery and the protagonist's dark past by themselves do not do much to dispel the realism of the story. The idea that a baker would disarm the dangerous situation of being threatened by some youths by sitting them down and listening to music together is unique but not in fact unbelievable. This build-up of realistic elements then clashes directly with the sudden introduction of the curse and does not produce a coherent genre result. Is it a crime thriller? Those do not usually contain real curses, but perhaps they contain imagined ones. However, it is usually part of a crime thriller that the detective protagonist questions the curse. This protagonist does not, and he is also the perpetrator of the crime. Is it a horror? Possibly, but listening to Wagner music is not an expected element of a horror story, and while the extreme hunger felt by the protagonists might be closer it is never portrayed with much urgency. Neither character is shown struggling to speak or think normally despite the hunger. Is it a comedy? When the idea of the baker tricking the boys into listening to classical music is first brought up there are strong indications that the result could be humorous. The introduction of the curse similarly could go in this direction, but it does not, offering no punchlines or sudden subversions. Instead both characters rather matter-of-factly agree on the curse and the new robbery and get in their car to go and find a victim. The decision to rob a McDonald's and the robbery itself again have many comedic aspects to them, but the overall plot and its conclusion are not humorous.

Based on the above, I would suggest that it is possible for readers of the original text as presented in the control version of the experiment to have used the textual plausibility to mitigate errors elicited by the text, but at a processing cost. The additional elements of textual plausibility introduced for the curse, the attack on McDonald's and the explanation of how the husband's past behaviour caused all of this would need to be understood in the first reading, and correctly extracted from the prior predictive models within a reader's working memory. This is possibly made more difficult by the fact that the mitigated predictive model still is not easily integrated into the global model as there is no further information nor any criteria of plausibility one could draw upon from any previous genre knowledge. The result would be as observed in the data for the experimental control: readers must carefully read the story and utilise their working memory to piece together elements of textual plausibility while not being able to speed up the process by relying on global model integration. Knowing this, it would be feasible to test this process more closely by selecting a text and manipulating it to have a number of impossible examples of

textual plausibility, with one thread of an actually plausible narrative embedded in it, and testing if and how readers engage with the actual plausibility.

For the participants reading the manipulated versions, this effect was modulated by the increased disruption of the manipulations. Within the description of the curse and the failed robbery, a number of manipulations confuse the exact time when it occurred, the protagonist's belief about whether or not he did the right thing to accept bread in return for listening to the Wagner album, and the location of the bread. In Version 2, the wife's statement about needing to rob another bakery immediately is manipulated to say "Before you're still hungry. You have to finish what you left unfinished." While the overall textual plausibility is perhaps still intact by virtue of her statement that they must rob another bakery, the subtle effect of these two manipulations may be that it is now entirely impossible to follow her logic as to why. In the first version a reader's predictive model would have to reconcile her statement that the curse must be intensified with suddenly speaking about lifting it, while in the second her insistence on robbing another bakery before they are hungry conflicts with the reader's knowledge that they already are. The textual plausibility is broken because in order to overcome our real world knowledge and fully enter a state of suspended disbelief a text must present textual plausibility without contradiction. The manipulated text no longer offers this, and as a result it is more difficult to reconcile the conflicting statements.

The increased difficulty in resolving these passages as giving a textual plausibility lead to two potential outcomes. It would be possible that readers work harder to integrate the new information with the prior predictive models in their memory and overcome the conflict there as well as the conflict with the global model by adding inferences and expectations of future information which will explain things. This explanation would however have to assume that readers either realize the text has been manipulated, or that they willingly choose to ignore part of the actual textual information in favour of the inferences and further expectations. Or, readers simply do not reconcile any such conflicts, do not gain any textual plausibility in their reading process, and instead substitute their own plausibility criteria from their global model: nothing in the story makes sense. They expect that they will encounter conflicts such as this by this point, as there have already been other manipulations which throw key story elements into question. The passages in question are less likely to lead to higher processing, since readers have no particular reason not to take the narrative at face value, and the consistency of the difficulty makes it clear that it will be difficult if not impossible to logically reconcile all of the contradictions. The predictive

models contain the expectation that the text will be surreal, illogical and therefore textual inconsistencies no longer trigger error signals which propagate up to the level of global model integration. It follows that this readerly plausibility overrides all others, leading to no additional errors being elicited by any of the following manipulations, which can also account for the reading speed being faster in the manipulated versions. The sections still need to integrate into the reader's global model at some stage however, and still deliver an overall narrative which does not cause readers to entirely give up or stop reading, and as the resulting data as outlined in sections 4.2 and 4.3 shows, participants did speed up, indicating that they expended less cognitive effort, while overtly reporting generally no more perceived difficulty than readers of the control. The difference was that those reading the control were able to activate more specific knowledge from working memory for some sections than those reading the experimental manipulations, and taking contextual plausibility on board within predictive model theory can help us to explain why.

In this section I have introduced and discussed the principle of contextual plausibility. I have shown how key cues within a text give rise to causal inferences for readers, and how these overlap the principle of minimal departure to create a kind of plausibility based on the textual input which is only modified by real world constraints after the cues have been resolved. I have shown how predictive model analysis can show these cues forming plausible predictions to explain input and resolve possible errors, even if the explanations themselves would not hold for real world observations. I have also argued how the experimental results I obtained show and support the finding that readers of the manipulated text would have been able to speed up reading times by utilizing reduced contextual plausibility to form overall predictions that the text is not plausible, leading to fewer errors, and less overall processing of the text overall, due to needing fewer new predictive models and error suppression steps.

This gives us a first part of the answer to our research question: using predictive model theory, what does it mean to understand a fictional text describing events which never happened and how does this happen in a typical reading process? It means that a reader forms bottom-up prediction models while also actively modulating their expectations and global model predictions in order to form closed predictive models which can satisfy the predictions and silence all error signals by supplying at least one level of causal explanation. The stage at which we have either suppressed all errors or formed a series of predictive models with the fewest remaining errors as we are able to tolerate is the one at which we break off. "Understanding" here of course has many possible connotations. In terms of

understanding the English language for instance, we would be able to say a reader has understood as soon as all the inputs have been matched to their possible predicted meanings, regardless of whether the predictive model can successfully match global model explanations or if any errors remain. Reaching a kind of philosophical level of "understanding" in an epistemic sense would be difficult if not impossible to define, as I have discussed in Section 3.2, as it may well be possible for an individual to suppress all errors without satisfying the kind of meaning we would like to attribute to the term "understanding".

A key question which these results left open was: how would readers have reacted if they had been able to clearly identify the genre of this story? What answers might I have received had I asked them to identify it based on the first reading? In order to answer this question as well as to more fully analyse and describe how genre plausibility and global model integration work, I developed a theoretical account of genre plausibility and tested it in another experiment. In the next chapter I will introduce and present the second experiment which arose from these considerations as well as the results and conclusions that I was able to glean from it.

Chapter 5: An Empirical Test of Genre Expectations

5.1 Story Structure as Genre Plausibility

In this chapter I will present the second eye-tracking experiment I conducted for this thesis. In the first section I will begin by expanding on the theoretical considerations which led to the second experiment. I will begin by introducing the notion of genre plausibility, which followed on from the principle of textual plausibility as I have discussed it in the previous chapter. In section 5.2 I will contextualize my ideas of genre plausibility against the current literature of text processing and literary linguistics. In section 5.3 I will introduce my methodologies and motivations of the experiment itself and my expectations and predictions before running it. Finally, a full discussion of the results will follow in section 5.5. Together with the conclusion of chapter 4, this chapter will offer the second half of an answer to the final and most important research question of this thesis: using predictive model theory, what does it mean to understand a fictional text describing events which never happened and how does this happen in a typical reading process? (see Section 1.1).

With some good examples of literary situations which require knowledge of the genre of story they represent discussed in section 4.4, it is necessary to discuss how exactly genre might work from a predictive model theory perspective, and what kind of knowledge structures readers will have about different genre types. These may be found in either fictional or non-fictional texts and being fictional is a trait of far more types of narrative than just literature. Any explanations of literature that draw on fictionality as a defining feature will fall into the trap of no longer speaking about literature but a great many forms of narrative systems which heavily rely on fictional elements within their structure. I will speak of fictional literary texts in the following as this is my main area of interest, but this theory could just as well be applied to any non-fictional or non-literary text. Under fictional literature I understand any fictional text written for the primary purposes of entertainment and without any emphasis on any particular notions of quality or prestige – the category will contain anything from world famous classics to fringe works, in any genre. One defining feature of fictional texts, or fictional literature, is that many fictional literary works begin to fall within perceived families of stories, or genres. I will not be able to account for genre as it is researched in genre studies in great detail as that is far beyond the scope of this thesis, but I do believe that genre and the reader expectations stemming from genre structures

are very important in literary reading. Instead of dealing with the wealth of descriptive scholarship on genre structures and their evolution, I will look at genre from the purely subjective viewpoint, as a learned experiential structure based on the reading done by an individual. Based on the sheer breadth of fictional literature which exists, I do not believe that any reader except for perhaps some of the foremost researchers of genre truly have knowledge of all genres, or anything resembling complete knowledge of any single genre. Instead, readers will possess a quite idiosyncratic knowledge of genres based on the stories they have been exposed to or which are culturally relevant. As a result, the kind of expectation structure I will discuss will not be the same as that of a work of thorough genre studies. It will however allow me to be more inclusive, and allow applications to any kind of text rather than simply fictional literature. For now, I will focus on fictional literature and the phenomenon of literary genres, while making use a non-literary examples as well, to explore potential perceived boundaries between the genre structures.

In order to give a convincing account of this we must find a good way to describe the possible ways in which knowledge generally might be stored and re-used for future predictive models. As discussed in Section 1.5, one of the properties of situation models as envisioned by van Dijk and Kintsch (1983) is that when fully resolved, the situation model itself can be learned as a whole. We will not need to follow this rather strong claim through in the sense of learning them permanently, but it is clear within predictive model theory that resolved predictive models stay in working memory and in a more direct way as the prior activations which necessarily will carry over into a new neural activation state. This use of predictive models as priors has great explanatory power. What exactly we retain in the long term remains the question and I will need to briefly return to some concepts from script and schema theories as well as the basics of Predictive Coding before discussing genre knowledge. It will necessarily be from the angle of “genre expectations” because the final structure resembles a more fluid, contextual process of predicting and expecting certain outcomes based on the genre a reader believes the current text to be and the kind of tropes in play within this current text, rather than rigid structures. A certain kind of rigidity one might call ‘schematic’ will remain in the fact that certain expectations will be linked to others by extremely strong associations, leading to inevitably expecting one thing to follow another. These expectations form patterns which become easier to recognize the more often one is exposed to them. With a basic structure for such expectations in place, we will move on to consider actual genre and text structures in order to naturally work out

the patterns of groups of literary texts and identify how genre expectations naturally follow from being exposed to these patterns.

One of the most basic types of pattern we can describe are simple causal chains or action stories, which we learn and complete automatically; causing us to duck from a thrown stone for example (Turner 1998). The stories are based on the concept of image schemata: schematic patterns in the human perceptual and motor system (Turner 1998, see also section 3.3). These simple causal chains become embedded in language use as action chains: "Action chains represent fundamental schematic experiences gathered in early childhood (things exerting force on other things). The energy source (the agent represented in a clause) transfers force to the energy sink (the patient in a clause)" (Stockwell 2012, p. 72). In this manner early schemata of causation become reinforced through both physical experience, and language experience.[5] These lines of thought line up very well with the structures called recognition models in Predictive Coding. These are pre-existing, learned Hebbian connections which react to input immediately and give an interpretation of the possible causes of the inputs. In essence their workings appear to be the same as predictive models, with the exception that no error signal remains, meaning that no additional explanation for an input is required. The major difference between frame or schema theory and my own is that frames or schemas are generally portrayed as rigid once activated, while predictive models are not.

While Turner gives them the name "stories", they are of course not the same in terms of scope and complexity as literary stories. The key question of this section is how genre works and how stories play a part in this. There is a curiously circular structure in which stories are defined by what genre they belong to, while genres are defined by the body of stories they encompass. Readers do not however read genres, they read stories. Therefore I will turn to story and story schemata first, by answering what exactly it means to be a "story".

Sanford and Emmott discuss conditions for turning narrative into story by adding some well-known terminology within literary theory:

[5] This point stands regardless of taking the viewpoint that this is either a causality somehow correctly observed, or a causality which is merely supposed by the observer as I have argued in section 3.3. Either way, the correlations assumed by the observer are learned as associations for future predictions.

- Setting (establishing main character(s), location and time).
- Theme (consisting of a goal for the main character(s) to achieve, possibly preceded by some specified event(s) which may justify it).
- Plot (one or more episodes, in which actions are performed in an attempt to meet the goal or sub-goals; realizing these may be temporarily thwarted by events that block these realizations, possibly leading to further sub-goals and additional attempts).
- Resolution (the realization of some event that satisfies the main goal leading to a state – e.g., a satisfactory outcome). (Sanford and Emmott 2012, pp. 4-5)

Plot is the important addition to simple narrative structure on which everything hinges, but it is not clearly explained why it is important. The beginning of an explanation comes from a study performed by Brewer and Lichtenstein regarding story grammar and reader response. The aim of their study was to establish possible properties of a story schema. The researchers proposed that a distinction should be made between the event structure and discourse structure of a story (Brewer and Lichtenstein 1981: 365). They believed that on the basis of a certain configuration of either the event structure or the discourse structure which the author used, certain conditions had to be met for a narrative to be considered a story. In this case, they postulated that a story requires an affective response which sets it apart from a narrative:

> We propose that a story is a narrative in which information about events has been organized in the discourse structure to produce suspense and resolution, surprise and resolution, or curiosity and resolution. To produce suspense, the event structure must contain an initiating event with a potentially significant outcome. A significant outcome is an outcome with important consequences (good or bad) for one or more characters in the narrative. (Brewer and Lichtenstein 1981: 366)

They went on to produce a series of texts designed to test a set of hypotheses regarding their view of story schema. The basic assumption was that a text with a significant outcome which was withheld until the end of the text would produce suspense, followed by resolution of suspense. A text withholding the initiating event would produce surprise when the outcome was revealed. Texts which produced suspense or surprise in this way would be stories. Texts which either gave away the significant outcome early in the text, or alluded to a significant outcome and never provided it would not be stories. In their study, the texts featured three routine situations involving a man driving home from work, a man on a beach in Hawaii and a gardener raking leaves on a lawn. Using these as 'base

narratives', suspense was produced by adding a bomb installed in the car, an incoming tidal wave and a winning sweepstakes ticket dropped on the lawn by a passing car. Multiple versions of each text were created: suspense versions which either gave the beginning and the significant outcome in chronological order or gave the beginning with a hint at the significant outcome; misarranged suspense versions which gave the outcome away on the first page; no resolution versions which omitted the outcome completely; and finally surprise versions which omitted the initiating events but gave the significant outcome. Participants were asked to read these texts and on a seven-point scale rate their experienced suspense or surprise and to what extent the text they read was a story (Brewer and Lichtenstein 1981).

The results mostly confirmed the hypotheses: the base narratives and the suspense versions without resolution received the lowest story ratings. The suspense versions which hinted at the outcome did not receive significantly different ratings from the standard suspense versions. Curiously, the misarranged versions which gave the significant outcome away in the beginning received story ratings only just below the suspense versions, and even above the surprise versions of two of the texts. This lead them to conclude that theorists must make a distinction between simple event schemata, narrative schemata which structure events together and story schemata which deal with the additional discourse structure and cause affective responses in readers (Brewer and Lichtenstein 1981).

While the scope of this experiment is small, and the texts which were used comparatively short, we can gain a lot of useful information from the results. The findings of Brewer and Lichtenstein highlight a very important fact: narratives are not automatically stories; they become stories by virtue of a very specific kind of content. As long as this content is present in a story the actual order of events does not matter, and even the beginning and end structure may be disrupted. This suggests that there was a clear expectation structure in their participants' minds for what a story is, which was directly tied but not equal to their conception of what a narrative is. The relationship is asymmetrical: stories must at their core contain elements which would make a valid narrative, but simple narratives are not stories. A story must represent events, causally relate and order them, but it also has to contain something significant. The significant thing represents the tension of the plot. Fulfilling the expectation of this plot structure is clearly significant to the perceived nature

of a text as a story instead of a simple narrative, and this may be an important part of a story schema.

I believe that this satisfaction of whatever a story schema may be is in fact the satisfaction of contextual plausibility when a text is compared to the global model of a reader. Brewer and Lichtenstein chose remarkable additions to their texts in order to produce suspense, while offering no explanation as to why they should work or why readers should intuitively accept them. Going by general expectations of our real world, it is quite unexpected for there to suddenly be a bomb randomly installed in someone's car, or for a tidal wave to appear without any warning, or a winning sweepstake ticket to appear out of nowhere, but these are tropes which readers may have come across and scenarios which we can generally envisage happening even if highly unlikely and indeed suspenseful if experienced. Due to this, the participants of this experiment appeared to have had little problem accepting these circumstances and even rated them favourably. This fits well with the concept as outlined in section 3.2 that both real world and fictional knowledge come together in processing even if the final verdict of a reader is to consider the text fictional, ultimately accepting the factually impossible aspects of the text. As with any other check against the global model, the principally important thing is whether or not the new input can be successfully matched to what the global model has stored as past successful predictions about the world. The results would suggest that Brewer and Lichtenstein's participants were able to integrate these stories rather well. The material provided to them contained few elements of textual plausibility to draw on, and I have already argued that the participants are very unlikely to have relied on personal experience of these events. It follows that the stories must have contained some elements of genre plausibility, which participants could draw upon from within their global model in order to integrate and accept predictive models containing the fictional story elements.

The structure behind genre plausibility of this kind appears to be rigid in the sense that certain conditions have to be met for a story to be recognized in this way, but also quite open and flexible in order for a newly encountered story to be able to fit into the expectation and fulfil the conditions of the genre plausibility. The suggestion of van Dijk and Kintsch (1983) that this may occur via the learning of entire situation models is too rigid to be able to achieve this. Instead, learning parts of situation models, or predictive models, which are then flexibly recombined appears a better answer, and this is where it intersects with more traditional schema theory. Generally, schema theory appears better suited to

explaining a phenomenon like genre plausibility for another reason: predictive models and the structure of bottom-up processing which deals with the text in real-time rely on textual cues to fit predictions. A set of expectations initially defines what is going to count as an error, and how errors will be explained away, using a combination of previous knowledge and textual input. Genre plausibility is a good candidate for explaining what the initial expectations are, but because they are initial expectations as well as fully top-down expectations from the global model, they are not reactive in the same sense that predictive models and textual plausibility are. Genre plausibility is schematic in the sense that it is a relatively static set of expectations which are contained in our global model, ready to cancel out any incoming error by a matching stimulus. Once activated, these expectations form the top-down attitude which helps distinguish early on what will cause any remaining error signals within a story and what will be accepted. It modifies the suspension of disbelief into a more specific suspension of some subsets of disbeliefs.

Some important evidence for the difference in expectation and reaction was gained in a study by Zwaan (1994), who was investigating the effects of what he also called genre expectations on reading effects. The study was based on the same outline of the situation model as was incorporated into predictive model theory (see section 1.4) and aimed to find out whether readers allocated cognitive resources differently between surface structure, textbase and situation model based on expectations tied to specific types of text (Zwaan 1994). In this study, text segments from both literature and news stories were selected and presented to two groups of test subjects. Both groups received the same texts, with a slight amendment of the texts. One half of each group was explicitly told that all of their samples were examples of literature, while the other half was informed that they would be reading extracts from news reports (Zwaan 1994). The prediction of the researchers was that experienced readers would expect the plot structures of a literary text to be more indeterminate than those of a news story, and would thus focus more on the limited information given by the textbase before constructing a strong situation model, while the news story would cause a stronger focus on a situation model at the cost of text detail. In brief, subjects were asked to select whether statements given to them after reading each text segment were present in the text. These statements were in the form of verbatim copies of text sentences, paraphrases of original sentences, plausible inferences compatible with a situation model but not found verbatim, and implausible inferences as distractors. The results were then indexed for which level of representation they supported: yes to verbatim sentences supported a strong textbase representation, while yes to paraphrases

supported both textbase and situation model representation, and yes to plausible inferences supported situation model representation. The implausible inferences were a control supporting no representation at all. In both experiments, the reported results supported the hypothesis, showing stronger surface structure and textbase representation for the groups who believed to be reading literary extracts, and stronger focus on inferences and situation models from the groups believing they had read news stories (cf. Zwaan 1994). The conclusion we can draw from this is that readers clearly adapted their reading strategy and their error resolution around what type of text they believed to be reading, focussing more heavily on establishing logical and causal plausibility for non-fiction, while more readily retaining large parts of the text in working memory for fiction and allowing for more freedom regarding causal explanations and plausibility.

Given that the materials used the exact same texts, it is clear that this difference in attitude came about before the participants had begun reading, and that it was triggered by the fact they were informed of the text type beforehand. Of course, the term "genre" itself presents a difficulty as there are many meanings behind the term. In alignment with Swales (1990) I will differentiate between linguistic and literary genres. Zwaan has taken linguistic genre for his experiment, classifying the texts as non-fictional and fictional. I believe that the difference in reading strategies between genres from this point of view has been sufficiently argued. The crucial next step is to evaluate literary genres and the nature of genre plausibility created by them. Starting from Zwaan's results, I predict that within specific literary genres the difference in reading strategies will occur not only as strongly, but with yet more specific initial expectations and suspensions of specific beliefs. Fantasy stories are allowed to have fantastical tropes and magic, science fiction is allowed to have technology and advancements that are not possible, while elements which do not fit, such as a genuine ghost in a detective story are still disbelieved and cause errors. To return to the beginning of this section, and schema theory in general, I propose that the global model is structured in this schematic manner, which can best be characterized through a combination of theoretical considerations of script theory, Barsalou's (1999) perceptual symbols and van Dijk and Kintsch's (1983) macroproposition structures which can explain how genre plausibility works and how it is learned.

In this chapter I have introduced the notion of genre plausibility. I have argued that there is good reason to view genre in terms of learned patterns of plausibility within texts, whereby we naturally group texts in which certain fictional elements are allowed together. These

form schematic associations in the global model. I have argued that these differ from textual plausibility in the sense that rather than being developed in the moment and modulated by context, genre plausibility is predicted top-down, from the global model. It is entrenched or modified by encountering textual plausibility. I have discussed that I believe this to be the source of personal and idiosyncratic definitions of genre which individual readers might have, irrespective of scholarly definitions of genre in the wider sense of literature and the arts. I have also shown how this view fits well into existing schema and script theories, and into the event-indexing model, while supporting my own theory of predictive model processing.

In the next section, I will discuss the schematic nature of genre plausibility in more detail, and review the way in which textual plausibility and repeated encounters with texts become expectations through predictive model processing over time, and how these manifest in the global model. To do so we will return to Schank and Abelson (1977) and Barsalou (1999) once more, as well as to the work of Ryan (1991).

5.2 The Schematic Nature of Genre Plausibility

In this section I will delve further into the nature of genre plausibility and return to the literature on script and schema theory, specifically the notions of Schank and Abelson. Using their theory as well as traditional genre studies work of Swales, Hebbian learning and the core principles of Predictive Coding and predictive model theory as I have thus far discussed I will show how genre plausibility is encountered by readers, and how I believe specific aspects of textual plausibility are entrenched and form predictive models which later go on to become further entrenched in the global model as predictions which we might call a specific reader's definition of a given genre.

Schank and Abelson (1977) originally assumed their scripts (or schemata) to cover specific situations, such as being in a restaurant or some social interactions. This creates the problem of how some scripts can interact with each other, or become activated together and fit into a more complex overall structure. It must be possible to explain how an individual is able to understand new or unexpected situations for which no script has been learned yet. For this, they suggest a concept of a higher order which they call a plan. A plan describes a set of choices used to accomplish a goal, and connects scripts to each other in

order to accomplish this goal (cf. Schank and Abelson 2008: pp. 69-73). They define it as another type of script-like structure: "Thus, plans are where scripts come from. They compete for the same role in the understanding process, namely as explanations of sequences of actions that are intended to achieve a goal. The difference is that scripts are specific and plans are general" (Schank and Abelson 2008, p. 72). This explanation of plans suffers from a weakness of distinction between a plan and a script, while at the same time suggesting that they are two separate entities within the cognitive system which is unlikely (cf. Strasen 2008). It does however offer a good starting point for appreciating the concept of genre. I will follow Swales, who defines it as follows:

> A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. […] In addition to purpose, exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience. If all high probability expectations are realized, the exemplar will be viewed as prototypical by the parent discourse community.
> (Swales 1990, p. 58)

For the present purpose, across both linguistic and literary genre, little needs to be said for the first two statements. Texts are clearly communicative events, in which one member is absent or distant. The purpose of a literary text could be to entertain or to bring across specific intended meaning by the author. As I am concerned with the understanding of text from a reader's point of view, I will not go much further than to acknowledge that authors may have such or other intentions, but also point out that often a reader will assume intentions on the part of the author which may be entirely false. Authors will use choices of style, specific words and tropes in order to make clear what the intended genre is, while readers will use their background knowledge to identify these choices. Readers may equally simply consider the purpose of such texts to be whatever they desire from a text at that moment, be it entertainment, education or something else.

Of note is Swale's view of genres as exemplars, to which texts can conform more or less, depending on how many existing expectations of the discourse community they realize. This viewpoint is informed by the fact that within literary genres, and in particular literary criticism, there has long been a discussion surrounding the importance of genre transgression and transcendence. Literary works have been praised for not conforming perfectly to genres, or for innovating genre structures. Swales here turns to the view that genre as a concept can only be transgressed if there is regulation and a normal structure in

the first place (Swales 1990, p. 36). This later leads him to the further difficulty that within prototype and exemplar theories it is by definition not easy to decide which attributes make members of a category more or less prototypical. Turning to Wittgenstein, both Swales and other theorists lean on the definition of family resemblance, in which Wittgenstein suggests that certain categories have no finite pool of attributes which every member must have, but that they are formed by chains of category members sharing certain attributes with another member, which in turn shares other attributes with other members and so on (Wittgenstein, 2006: pp. 277-281). There have been many scholars who have offered the counter-argument that this in theory allows one to see family resemblance between all objects in the universe. I believe this to be a poor counter-argument as the fact is that all objects in the universe share common traits no matter which sorting criteria we use; they all exist, they are made of matter, interact with energy etc. The concept of family resemblance combines well with prototypes as long as we remember that it is about perceived similarity just as much as actual similarity, and that we are not talking about all possible attributes in the universe, but those an observer knows. A genre is a prototype based on an individual's own experience and knowledge and how well they believe a new object to conform to the body of others they know. That is, precisely as Swales says, if the new text confirms that individual's expectation of what a text from genre *n* should be like, then that is its genre, for that individual. This interpretation fits in very well with both predictive model theory, and with schema theory. What I wish to add to this is how genre structures are learned and updated.

The difficulty in defining genre and genre learning while at the same time acknowledging that many texts innovate or transgress the boundaries of the actual genre they belong to also leads to a difficulty for schemata and supposedly rigid mental structures. Schank and Abelson recognized it and suggested there exists a middle ground. There must be schemata which are as such separate from one another, yet can be activated together, contrasted with each other and lead to a combined output, that is, to a conclusion which could not have been gleaned from either of the schemata individually. Otherwise other phenomena such as blending (Turner 1998) or conceptual metaphor (Lakoff and Johnson 2003) would be difficult if not impossible cognitive tasks. To explain how this is possible, but also how the mechanisms of connected schemata lead to an overall network of schemata which could resemble what I have defined as the global model, containing genre plausibility criteria, this must be redefined in terms of predictive model theory and in more general terms. This is best illustrated by some examples. Let us consider an example once used by

Schank and Abelson (1977), regarding a “restaurant script”. They suggested going to a restaurant is a script, which contains slots for ordering food, sitting at a table and so on (see Section 1.2). I would like to suggest, as they did, that there is a lot more to this deceptively simple sounding example.

Behind the scenario of going to a restaurant, ordering some food, eating, then paying lies a very complex set of possible factors with potentially infinite variations of actual restaurant visits. Just to name a few factors: is it a restaurant where one finds one’s own table, or is one seated? Does one order at the table or the bar? What kind of food is served, at what time of day, of a specific ethnicity, several, or none? Does one pay before eating, after eating, for individual items? Given the individual sequence of the above, coupled with where exactly any given restaurant-goer might be seated, what they order and how they feel about it there is an uncountable number of possible combinations. No single script, schema or pre-built predictive model could possibly hope to account for them all. Instead, rather organically a skeletal schematic structure will emerge via repeated predictive models and become encoded into the global model through the shared features of a restaurant visit. Both the original scripts and the later plans of Schank and Abelson (1977; 2008) can be reconciled with the predictive model theory view if we consider the nature of perceived events and event strings.

The term “event” is as difficult to define as “situation” as I have previously discussed in section 2.4, but also plays an important role in thought and reasoning. Let us assume that we are reading about a series of restaurant visits in stories, although the following argument could just as well apply to actual visits to real establishments. A useful definition of events for the purposes of literary theory is the following:

> Events are perfective processes leading to a change in truth value of at least one stative proposition. State propositions fall into two categories: some express inalienable properties, and retain the same value throughout the narrative (x was a wolf, y was the daughter of a king), while others present the potential of alternating several times between truth and falsity. (Ryan 1991: 124)

This event structure could also be called an image schema according to Turner (1998), or an action chain as discussed by Stockwell (2012). Sweetser (1991) similarly suggests underlying cognitive structures based on physical experience. Over time, certain aspects of specific events are learned via repetition and become part of the global model. Not every part of a perceived event will be stored, as previously discussed in section 3.3, following Barsalou’s

concept of perceptual symbols (1999). There are two reasons for this: it is biologically unfeasible to store the vast amounts of data required for recollection and learning every last detail perceived, and it is in fact pragmatically useful to only store key aspects in order to make the retained details maximally useful for future predictions. To explain the second reason, let us consider some examples of different possible restaurant visits.

1) A hungry student visits a café that sells hot food. The food is ordered at the bar and paid for in advance. It arrives at the table the student picked after about 10 minutes, and the student is satisfied with the food.
2) A middle aged couple celebrate an anniversary at a special Italian place downtown. It is very busy when they arrive. They wait around 20 minutes to be seated by a member of restaurant staff. Waiting staff take a drinks order, followed by an order for starters, and main course. Starters arrive after half an hour and are eaten and cleared away before the food arrives some 20 minutes later. After they have eaten they linger a while, and eventually ask for the bill, which is substantial. Feeling satisfied, they pay and resolve to come back the next year.
3) An office worker joins some colleagues for lunch at a sushi restaurant. The worker is very confused by the central kitchen, and the rotating band on which new sushi is placed. The colleagues explain that food is not ordered traditionally. The sushi chef in charge makes whatever he feels to be appropriate, but will take requests from customers, who then take what they want from the band. The office worker does not enjoy the food but stays polite and pays for the eaten food at the end, resolving never to come back.
4) A hungry construction worker orders some food from a favourite takeaway place through a website. The payment is taken immediately by card, and the order sent ahead to the restaurant. The worker stops by the place on the way home, where it is already packed up and waiting to be eaten at home. The worker is very happy with the food but wishes it had stayed a bit warmer.

Each of these examples contains something important about restaurants, but also information that was not necessary in order to understand either the example or what a restaurant is. Even from just these four, it would be possible to form a reasonable schematic expectation for our global model which could deal with a large variety of new examples. Let us assume that we begin with example number one, and this is perhaps the

first time we have encountered a restaurant situation. Each item of information might be learned, which we then attempt to re-use. Going forward, we might expect that restaurant visitors are hungry students, food is ordered at a bar, then brought to a table picked out by the student and is generally satisfying. While we may not expect a single such experience to become ingrained in a global model and future predictive models, it could be held within an individual's memory with this level of detail. How useful is this now memorized schema for other restaurant scenarios? Not as useful as we might expect. Let us say that we have fully processed example 1) and represented it as a full predictive model n. We now encounter example 2). and an error as shown in Figure 21. We are attempting to apply a prediction of the student as the restaurant goer, or main protagonist of the situation, but of course there is no student in this new scenario, so the model must adapt.

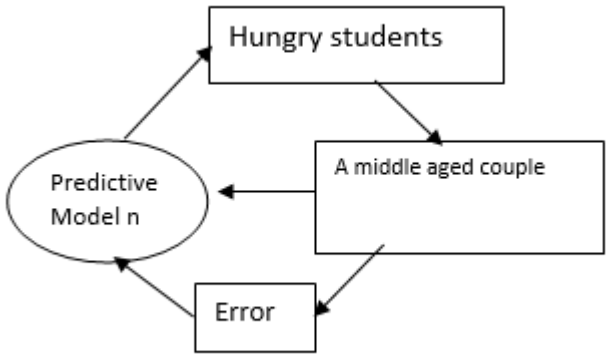

Figure 21: An error encountered by the mismatch of expected protagonist and actually encountered protagonist

In terms of the event structure suggested by Ryan, this is also an event, both in the sense that a prediction is confirmed as false, and in the information that we have a new property for the middle aged couple as restaurant goers. Stereotypical character roles may be learned in this way, based on the actions and general agency allowed certain entities or characters, and what types of scenarios we reasonably expect them to appear within. Next we receive new information which was not contained in the original predictive model. The couple encounter a busy restaurant, and must wait for a table they have not picked. This elicits errors for several reasons: there is nothing in our prediction based on 1) regarding whether restaurants are busy or not, so this information is not covered and must be sent up the hierarchy as an error for a new prediction. The couple also do not get to choose their table, but are directed by a member of staff. This is an introduction of a new agent as well

as a different level of agency for the restaurant goers. Finally of course the table is mentioned in a much earlier sequence than in example 1). An entirely new prediction must be formed around this circumstance, as so far in fact nothing from our first predictive model, n, has actually been helpful in predicting the new situation. The new predictive model, n+1, must incorporate an explanation of the new information such that apparently middle aged couple can also go to restaurants, and that restaurant goers cannot always pick their own tables, as shown in Figure 22.

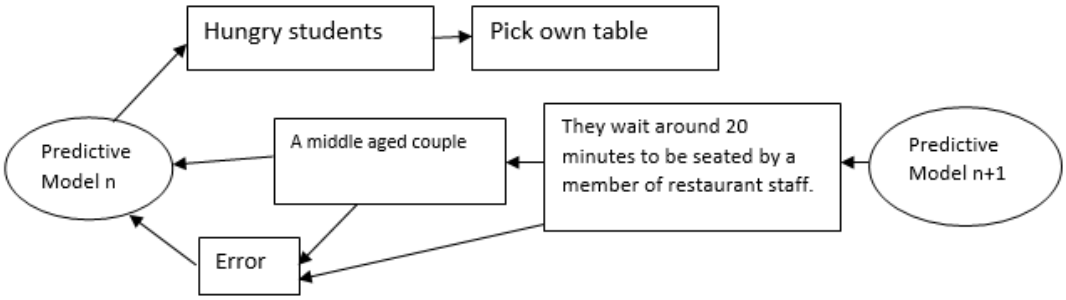

Figure 22: A new predictive model n+1 must be represented to explain the new input, which the existing predictive model could not account for.

The new prediction is favourable because there is no internal plausibility issue. There is no logical contradiction between new and old information, as the first example never contained the suggestion that only students can go to restaurants, and that one may only pick their own table etc. Given that the information we can receive is finite, we must accept our current explanation until it is falsified. For the same reason, this cannot be said to be a case of textual plausibility, as there is no explanation or fictional causal relationship within the example. The number of acceptable details for an otherwise similar predictive model is being extended. This scenario concludes with the new information that drinks, starters and main meals can be ordered separately, and may take additional time to arrive, and finally that the bill can be paid at the end. To contrast it with Swale's definition of genre, this example has transgressed the previous boundaries of the *restaurant* genre, but not formed an entirely new one as we are still the same communicative population and both examples had the same communicative purpose and fulfilled the expectations of eating at a restaurant.

As a result of having processed both of these examples, an already more nuanced prediction is possible for dealing with example 3). We now realize that more types of people can be restaurant goers, that ordering can work in two different ways, tables can be

picked by staff or the visitors and payment can be taken before or after. Still, for example 3) new information must be taken into account. Another new agent, the office worker, is introduced and once again the restaurant works quite differently. This time instead of tables, there is a large circular bar with a revolving band. Each of these things will result in errors and new predictive models being formed. The same will happen for any new example or specific information. When considering how the global model needs to be updated in order to best deal with any new situations, we can see that most of the specific knowledge regarding for example who specifically the restaurant visitors are is not useful in order to successfully predict most situations. Instead I propose that only key details are needed, but these key details form the core of any predictions for similar types of scenario going forward, and in fact supply the criteria for grouping experiences under a combined label such as "restaurant visits". The one constant across the first three examples is that three different kinds of actors who are all humans physically go to a location, the restaurant, where they have food prepared by another which they pay for.

To put this in terms of Ryan's event structure, we have identified a few conditions which seem necessary to recognize a restaurant situation: 1. The actors must be human, as an inalienable property. 2. The actors must go to the location of the restaurant, which then becomes a true property. 3. The actors must receive food, which becomes a true property. 4. The actors must pay for the food, which becomes a true property. By using this event structure as a framework integrated into our global model, we could begin to successfully identify many different restaurant scenarios, involving any number of other details, by only focussing on predictions requiring these four properties to be fulfilled. That means, regardless of what other information is contained in a description, a predictive model in which these four properties become true will always be successfully integrated with the global model. Because this is the end point of any processing, this contextual plausibility will override any other logical concerns with the situation, and it will be accepted as a restaurant situation, even if there are violations of plausibility regarding other details. This will allow us to learn this new situation for future reference. As stated above, this is not textual plausibility which only occurs at the predictive model level and is taken from actual in-text descriptions. The scenarios contained no explanation of causality, or why the restaurants or restaurant-goers behaved the way that they did. It is a genre plausibility which has been learned by the overlap of repeated scenarios, which has become a causal relationship not because this is ever said explicitly but since it has been used by an individual's brain in order to explain the successive inputs and become part of their global

model. Because of the lack of determinate requirements except for the categorical properties, this kind of plausibility structure is perfectly able to handle any number of new situations which fit into the pattern, regardless of what other information is contained in each specific instance.

What was referred to as a plan by Schank and Abelson is the desired outcome of such sequences of events: having eaten. Ultimately all scenarios are about different protagonists trying to get food prepared by someone else and somehow fulfilling the requirements for it. This can be conceptualized as an ending, or a climax and it also satisfies the criteria of plot as described by Sanford and Emmott as well as Brewer and Lichtenstein. Specific endings and climax structures may be confined to individual examples, but the underlying event structures of what may constitute an ending or a climax of an event sequence could be learned as a schema structure within the global model. If payment comes early in the sequence for example, the eating itself may form the ending, while in a scenario like example 2) where more actions follow, the eating would form a climax. This structure can then have interesting effects when coming across example 4). To some readers, the take-away food situation may actually fall outside of their expectations, as the food is not consumed inside the actual restaurant. For others, who are very used to this kind of scenario it may be a completely non-problematic instance of a restaurant situation. For each individual, the first experiences define something akin to a prototype, one that is very overdetermined and contains too much detail. Over time, details that are not useful for future predictions fall away to be replaced by more schematic expectations as above, but which expectations exactly is a matter both of individual experience and of cultural overlap.

Some scenarios will merely satisfy the ending conditions of this restaurant schema; others become associated with specific event structures and a cultural context. These others, together with the schemata of certain characters and settings, form what we could call *genre*. For example, “Italian restaurant” is something which likely means a very distinct thing to many people, while “Chinese restaurant” or “Steakhouse” mean very different things. They are formed by repeated visits to certain establishments that share key traits with each other and with what is generally acceptable across a culture. They form the type of prototypical structure which fulfils and completes the definition of genre put forward by Swales: genre transgression is only transgression at the local and individual level, but over time forms part of the genre, which is itself a fluid structure of knowledge and expectations within an individual’s global model of the world. Sometimes texts may fall too far outside of

the boundaries for one reader, while being acceptable to another. The conclusion that I would draw from this view, as opposed to the classical literary view of genre is that there are several types of genre: individual genres, as readers experience them, and cultural genres which are shared and negotiated by a population. I do not have the space or data to further discuss the idea of cultural genre but I believe it would make a fascinating area of further study.

The individual genre schemata are encountered many times in different individual texts, whether actually eating at certain places, or coming across them in conversations with people, depictions in the media or descriptions in various literatures. While it may not be possible to give an accurate account of an exact genre schema present within an actual individual's global model, one way to portray what a general structure of genre expectations and expectations about stories look like is to turn to the structures identified by literary theorists. Much of research is about finding the patterns which are important, wherein the patterns themselves are sometimes determined by the patterns humans impose on the world in the first place. A very good example is the diagram of the narrative communication model across two levels as developed by Wenzel (2004), shown in Figure 23. Wenzel considers the elements shown in the diagram to be the most important elements within the discourse structure of any narrative text. I not only agree, but also believe that this identifies more than discourse elements: it identifies the most important aspects of a narrative which will be expected by a reader and which in turn form a reader's predictions about what kind of a narrative is encountered.

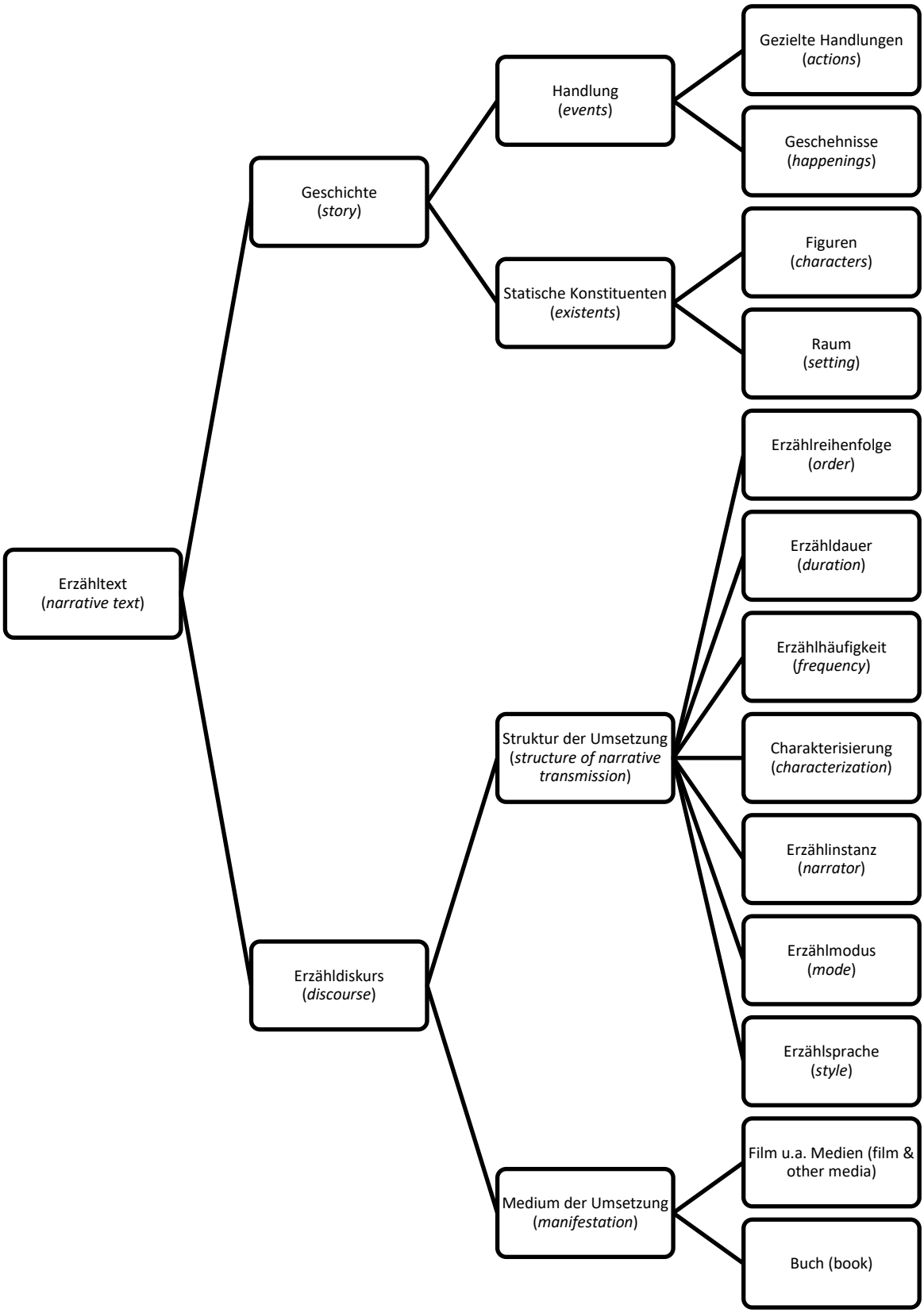

Figure 23: Zweiebenenmodell mit ausgestalteter Diskursebene. The two-level model with developed discourse level. From Wenzel 2004: 15.

The plausibility of a general story schema, which in a reading situation provides what I call contextual plausibility, can offer an explanation for expectations about what a story is and how it should be told. The sections of Figure 23 above can serve as a good starting point for identifying possible schematic nodes or "slots" to use Schank and Abelson's (1977) terminology (see Section 1.2). While many of these slots will depend on personal experiences as I have discussed in Section 5.1 and in the examples 1) – 4) above, some of them are also cultural and thus more widespread, and some will be universal. For example, in written literature the "manifestation" slot will always contain the requirement that it be written. Some will also satisfy the "book" slot, which also brings specific predictions and schema requirements with it. For example, books must be read one page at a time, in a specific order. More importantly, books are subject to specific story structures which are not shared with film or other media: they are organized into chapters, and within chapters into paragraphs and distinct written sections which can be identified not just by content but by their graphological appearance, spacing, and of course punctuation. The book slot also places restrictions on or opens up possibilities for other slots. "Narration" has more possibilities in combination with books than some other media for instance. Books can offer true first-person narration, which films struggle to do and bridge with devices such as voice-overs, which still cannot convey the internal thought and feeling perspective offered by a novel. "Duration" meanwhile is more indirect for written texts, while being very specific for film or episodic series: every film or episode has an exact duration. It is hard to quantify how long it takes to read a book, so instead a more indirect measurement of page length is generally used. While publishers and authors generally have very clear concepts of what page lengths are suitable, the actual enjoyment of certain novel lengths and reading duration will be entirely subjective and vary across readers.

The most variation across genre structure lies within the "story" slot and sub-slots of Figure 23. While certain story-structures will overlap across genres, some of the specific constituents, events and character archetypes will be unique and highly idiosyncratic to certain genres. Some of these are established via repeat readings of similar stories, leading to very strong expectations of character archetypes and event structures. In the same manner as any real experience, literature contains sequences of events, climaxes and outcomes. Just like in the restaurant examples, while reading an individual is always attempting to match the current input to their global model efficiently, leading to certain expectation structures. A reader might be faced with many stories in which a murder is committed. In some cases this is simply an event amongst others within the story which has

an otherwise quite funny or entertaining outcome. In many cases however, the murder will be accompanied by a very special type of character who investigates this circumstance. With it comes a very special type of character who has committed the murder. A third type of very specific character fleshes out the story by being the potential murderer, until it is proven otherwise. As a reader encounters stories which have different outcomes, different murders, perpetrators and investigators, the specific patterns of plot which correspond to the conflict between murderer and investigator and the interaction between them overlap and begin to form a network of expectations. A reader comes to predict that when one of the important character archetypes is identified, another which is closely related to the role must also be identified. It becomes obvious that the suspect identified at the beginning of the story will rarely turn out to have actually done it. The expected outcome is that of course the investigator will find the right one, but at the very end of the story, and so on. This forms a schematic baseline of detective novels in the global model, which in turn incorporates more detail the more overlap a reader encounters within different stories of this type. The attributes of what investigator characters are like can become heavily stereotyped, causing certain traits associated with them to occur time and time again. They are intelligent, strong willed, either private detectives or police detectives. A popular archetype sees the police detectives often being portrayed as unconventional and at odds with their own colleagues, or the private detective being a rogue equally at odds with the local police, in both cases representing the police in general as incompetent while only the protagonist is competent. Entire tropes and internet jokes exist about the overused occurrence of the police protagonist being suspended from duty halfway into a plot only to redeem themselves by stubbornly continuing to work. The cases they have to deal with also follow consistent plot patterns. In the recurring novels of master detective Sherlock Holmes by Arthur Conan Doyle, they are always murders of a mysterious nature which the police are not quite capable of solving. Holmes inevitably takes on the case, before disappearing and reappearing for the climax of the story where he presents the solution (cf. Busse 2004). In similar stories, the plot arches may develop differently, and the main character is of course not Sherlock Holmes, but important parallels exist. The differing levels of character and event archetypes can further interact with the medium in interesting ways. In this particular instance I am concerned with texts, which have almost total freedom to describe events and actions linguistically and require more on the part of a reader to imagine or follow such events. Movies are more adept at directly showing how such events happen but subject to different constraints given what is possible to film or increasingly to simulate using computer generated graphics, but also in the sense that it is impossible to film an

individual's thoughts. This leads to film makers using different techniques including additional narration and voice overs which changes the style and tone of the transmission. *In each case, the degree of detail of any such genre schema present within any given reader's global model depends on the amount of direct and indirect exposure the reader has had to stories of this type. For some, crime thrillers and detective stories may form a fuzzy whole with little distinction, while for others they may be highly distinct and unique schemata*.

In this section I have discussed how genre plausibility forms by looking at the possible origins of a schema for a specific situation. Using the restaurant schema as an example, I have shown how specific parts of predictive models can become entrenched and form useful additions to the global model. I have also argued that these are likely to be far less detailed than full predictive models, but that this forms a pragmatic advantage. A schema which is too specific fails to be useful for predicting other situations than the one it describes is not useful and creates more errors than resolves. Instead, these schematic predictions must necessarily be less determined and contain open slots with only baselines requirements in order to account for as many variations on the baseline scenario as possible. I have also discussed some specific examples to show how these can naturally form and what shape such schematic genre plausibility may take based on a few inputs.

To build upon this, I decided to design a second eye-tracking study which would investigate how real readers deal with established genre tropes in an experimental setting, and to what a degree prior knowledge of a text and its pre-conceived genre identity mattered. In the next section I will introduce this experiment in detail, and present the results.

5.3 Experiment: Genre plausibility and expectations

In this section I will introduce the aim of my second empirical experiment performed while researching this thesis. I will discuss the motivations and results from my previous experiment which led to it, and the process of the new design.

The aim of this experiment was to investigate how reader's prior knowledge of genre plausibility influenced their reading of a text. In deciding how to design it, I considered the important aspects of several other studies which had been instrumental to predictive

model theory so far, in particular Zwaan (1994) and Brewer and Lichtenstein (1981) as well as once again the event indexing model (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998). The experiment of Brewer and Lichtenstein (see also Section 5.1) focussed on how distinctly readers perceived certain texts to be stories, by introducing different kinds of plot. Based on this it was clear that certain types of plots and genre plausibility lead to this perception being enhanced while a lack thereof leads to a more difficult perception. I intended to once again work with real literary texts rather than constructing my own examples, so as to gain data based on real texts and to avoid the pitfalls of testing texts based on my idiosyncratic interpretations of popular genre. This meant it was impossible to use texts which contained no genre elements, but I also needed a contrast between texts so as to be able to measure differences between their receptions by readers in an experimental setting. Zwaan's experiment on genre expectations (1994) offered an interesting alternative: some participants were told that the same texts were either fictional or real newspaper articles, and this influenced their perception of these texts through the immediate prior knowledge of the text type. This went on to also inform the event indexing model of Zwaan et al. (Zwaan, Langston, & Graesser, 1995; Zwaan, Graesser, & Magliano, 1995; Zwaan, Radvansky, Hilliard, & Curiel, 1998) and the assumptions that increased event overlap leads to better and easier processing. As my first experiment showed, affecting event overlap does cause differences in reading behaviour, but does not necessarily lead to reading speed increases but rather modulation of predictions. To combine these assumptions as well as my theoretical position towards genre plausibility established so far in this chapter, I predicted that there should be a noticeable effect on reading and on processing speed if participants were made aware of the genre of a text and had their expectations regarding that genre fulfilled or conversely disappointed. I therefore set out to design the study in such a way that participants would read a number of extracts from specific literary genres and be informed beforehand of the genre of these extracts. One version would tell them the correct genre and another an incorrect one, in the hopes of seeing a noticeable difference in reading speed between these versions, caused by either a successful or unsuccessful overlap between expectation and text. I also wanted to avoid participants using a strategy of not engaging with difficulties in the text, hoping that the very genre-specific nature of the texts selected together with the direct exposition of the genre would not allow participants to freely modulate their predictive model expectations as easily. Therefore, even if they did choose to opt for an adaptive strategy of forsaking both their expectations of genre plausibility and the genre information of the experiment, there

should be a measurable impact of the cognitive effort to do so. I decided to test five genres: *Romance*, *Crime Fiction*, *Fantasy*, *Science-Fiction* and *Horror*. In the following sections genre labels will be capitalised and italicised to make clear when I am referring to the tested genre in question. All five are established culturally in the UK and are represented as categories within book stores and online discussion in places such as goodreads.com and similar book review sites. This made it exceedingly likely that participants would have awareness of and some prior experience with these five genres, without me needing to find a more strict theoretical rationale for each genre. It also made it likely that participants would have some ability to recognise whether or not a particular text should belong to each genre.

Participants

Overall 34 participants were included in the analysis. A further six participants participated but were discarded - two participants were discarded due to a fault in the initial experiment design, another three due to bad recording and skim reading of large parts of texts and one dataset was lost due to a technical fault overwriting a previous recorded dataset All participants were recruited from the module Language and Context, a required first year Bachelor Degree module providing a broad introduction to linguistics, within the University of Nottingham's School of English, and received course credit for their participation. Participants were all native speakers of English.

Materials

In order to design a successful eye-tracking experiment with a limited duration to avoid fatigue in participants it was necessary to limit the overall genres tested. Taking into consideration a realistic number of text portions to use for an eye-tracking experiment and to give each tested genre enough material, I decided on five genres to be tested. To still ensure a thorough investigation, I also decided to use two examples for each. As a result the eye-tracking portion was made up of altogether ten extracts, with two representing each genre. The extracts were taken from real texts found on the Amazon Kindle shop (amazon.co.uk). Each text was selected because it was branded and sold as belonging to the genre being tested. The exact sources were:

Romance:

Longley, Barbara (2016). *What You Do to Me* (The Haneys Book 1) Montlake Romance. Kindle Edition. (Kindle Locations 755-792)

Liasson, Miranda (2016). *Can't Stop Loving You*. Montlake Romance. Kindle Edition. (pp. 181-182)

Crime Fiction:

Croft, Adam (2011). *Too Close For Comfort* (Knight & Culverhouse Book 1). Circlehouse. Kindle Edition. (pp. 65-66).

Robson, Roy (2015). *London Large - Blood on the Streets* (London Large Hard-Boiled Crime Series). London Large Publishing. Kindle Edition. (p. 86)

Fantasy:

Sanderson, Brandon (2016). *Arcanum Unbounded: The Cosmere Collection*. Orion. Kindle Edition. (Kindle Locations 1576-1586)

Hobb, Robin (2009). *Dragon Keeper* (The Rain Wild Chronicles, Book 1). HarperCollins Publishers. Kindle Edition. (p. 147)

Science-Fiction:

Dembski-Bowden, Aaron (2016). *Night Lords: The Omnibus* (Warhammer 40,000). The Black Library. Kindle Edition. (Kindle Locations 8346-8356)

Tchaikovsky, Adrian (2015). *Children of Time*. Pan Macmillan. Kindle Edition. (pp. 491-492).

Horror:

Wood, Rick (2016). *I Have the Sight* (EDWARD KING Book 1). Rick Wood Publishing. Kindle Edition. (pp. 49-50)

Nevill, Adam (2011). *The Ritual*. Pan Macmillan. Kindle Edition. (pp. 75-76)

From each of these novels, an extract of 450-550 words was selected and split into exactly ten sections to be used in the eye-tracking experiment for optimal reading. I selected them by searching the texts for examples which would make sense isolated from their respective texts with only a brief introduction, and which contained sufficient markers of the genre they represented. Each extract was preceded by a specific title screen in which I provided a brief summarized context for the extract itself, and which importantly stated to participants what the literary genre of the novel was. Every such title screen began with the sentence

“The following extract was taken from a much longer X novel” in which X was replaced by the appropriate genre. In version 1, the control condition, participants were told the correct genre. In the experimental condition, version 2, the genre labels were changed while the actual text extracts remained in the same order as version 1. I chose this method for this experiment in order to avoid some sections being labelled correctly, as I was interested in how many participants would correctly recognize all labels being incorrect. For future iterations of the experiment, it would be of interest to fully randomise labels instead. In order to provide a reasonable comparison only the actual genres included in this experiment were given as options, and the order of extracts was set. The final version of the labels given in order on both lists was as follows:

Version 1 (correct labels):	Version 2 (incorrect labels):
Romance	*Sci-Fi*
Horror	*Romance*
Sci-Fi	*Crime*
Romance	*Horror*
Horror	*Fantasy*
Fantasy	*Crime*
Sci-Fi	*Romance*
Crime	*Fantasy*
Fantasy	*Horror*
Crime	*Sci-Fi*

Following each extract, participants were asked a short comprehension question to ensure they were actively reading the extracts and engaged with the task.

Immediately following the eye-tracking, participants were given a survey to complete. In a series of questions the participants responded how many hours they spent reading each of the five genres present in the study, giving each one an individual response from one hour to five hours or more. They then indicated how much they felt they enjoyed each genre, individually rating them from strongly disliking to strongly liking, on a five-point scale. Following this, participants gave a favourite genre, or several if they felt uncertain. The final question asked participants if they felt that the genre labels for all read extracts had been accurate or not, with space for free-form answers. Participants were verbally encouraged to give examples or details if they could recall any. Some additional questions were asked for

the purposes of collecting data for future alternate analyses which are not included in this thesis. The entire survey is available in Appendix A.

Procedure

Recording was done using an SR Eye-Link 1000 attached to a recording PC and screen together with a laptop featuring an eye-tracker interface for the researcher. The eye-tracker recorded at 1000Hz, while participants were seated exactly 70cm from the display screen, using a chin rest to minimize head movements. Participants were informed of the purpose of the experiment and the overall procedure and health and safety concerns, then gave informed consent. Participants were instructed that they would be reading a series of novel extracts of various genres and asked to read it as if reading at home for pleasure. This set up took an average of 10 minutes.

Text was presented one section at a time, averaging approximately 50 words, triple spaced in size 16 font to ensure comfortable readability and more accurate recording. Participants read each section at their own pace and pressed a key on a keyboard in order to move on to the next screen. Questions asking about simple details from the story followed each extract and were inserted to ensure readers' attention and that the extracts were being read for comprehension. All comprehension questions were yes/no questions and participants pressed the n key for no or the y key for yes responses. The eye-tracking part of the study took an average of 40 minutes.

After each participant completed reading the entire text, they were given the pen and paper survey which took an average of 5 minutes. At the end, they were given the opportunity to ask questions and give general feedback.

Predictions

Given the strength and prevalence of prior knowledge and of expectations for specific genres as discussed in sections 5.1 and 5.2, I believed that participants should have trouble integrating the incorrect genre labels with the clearly identifiable extracts which follow. This integrative difficulty should result in participants needing to adopt alternate strategies of integrating the genre plausibility within each extract with the genre label given in each title screen. Since the extracts were in a alternating order it would be difficult for

participants to be able to see any pattern in the incorrect labelling, leading to a robust reading time increase compared to the correctly labelled condition. The hypothesis was:

If readers' global models contain schematic predictions of genre plausibility for the genres contained in the experiment, then being confronted with a correct genre label should lead to easier error suppression and global model integration, while incorrect genre labels should produce more errors and more difficulty with global model integration, leading to a difference in reading speed.

While in the previous experiment participants had modulated their predictions to reduce error and also reading speed for the experimental section, in this experiment the exposition of the genre should rule out the possibility of normalising the text as being less comprehensible, and lead to slower reading times for the experimental condition, as readers should need to expend further cognitive effort to suppress the errors of genre plausibility not being fulfilled by the extracts.

I also believed that, given the prevalence of the five genres tested, participants should, as members of the discourse community (Swales 1990), be quite accurate at indicating that the labels were incorrect.

In the next section, I will present the results of the study split into two different types of analysis that was run, followed by a breakdown of the results of the survey given post eye-tracking.

5.4 Results

Before analysis, the data was cleaned using the SR Data-Viewer software's built in 4 stage process, discarding fixation times below the threshold of 80ms and merging some smaller fixations within an interest area. All results were calculated using dwell times as reading times corrected for string length differences between versions. Dwell time was chosen as I wished to account for differences between the total reading times of sections as time spent by a participant before moving on to account for their full reading and comprehension process. This means total dwell times in each interest area, with each interest area being defined as one full word or contraction within each section, were divided by the number of characters contained in the area in order to account for the manipulated texts having slight differences in the amounts of letters to the control. This results in a value of reading time

per character, which can be compared between datasets even when the word or word length amounts of the samples slightly differ. The values were then log transformed, allowing for an accurate linear regression calculation to determine whether the differences between reading times across versions were statistically significant.

Analysis

For the analysis, the total reading times per character for both versions was measured, followed by the reading time per character for each of the genre conditions. This was done by separating the datasets into groups all of the text sections plus introduction belonging to each genre in version 1, our control, as well as their equivalent sections containing the incorrect genre label in the experimental version 2. For the purposes of reporting, the conditions are labelled by the genre the actual textual samples were from and "incorrect" to indicate the experimental condition in which they were mislabelled, and the genre the samples were from and "control" to indicate the control group which read them with correct labels. Reading times were in all cases slower, with more time spent per character in the incorrectly labelled version 2 texts. Reading times are reported in table 19.

Table 19: Reading times per character for experimental conditions vs control

	Mean	Standard Deviation
Incorrect labels total	73.6ms	56.1ms
Control total	70.3ms	56.5ms
Romance incorrect	72.3ms	55.7ms
Romance control	69.5ms	54.3ms
Crime incorrect	73.4ms	55.5ms
Crime control	70.6ms	59.4ms
Fantasy incorrect	72.7ms	54.5ms
Fantasy control	68.3ms	52.3ms
Science-fiction incorrect	76.9ms	62.2ms
Science-fiction control	73.3ms	62.6ms
Horror incorrect	72.8ms	52.2ms
Horror control	70.3ms	53.8ms

Linear regressions of the same datasets showed that across both datasets in their entirety, as well as for each genre individually, the log transformed reading speed per character difference was statistically significant. The results are reported by the original genre vs control, so for instance "Romance vs control" signifies the regression was run using the log reading times per character of the incorrectly labelled Romance segments in version 2 versus the same correctly labelled Romance segments in version 1 as dependent variable, with version as independent variable. Results are reported in table 20.

Table 20: Linear regression of log reading times per character by version

	Estimate	Std. Error	t value	Pr(>|t|)	
V1 vs V2 total	0.024080	0.001549	15.55	<2e-16	***
Romance vs control	0.019455	0.003414	5.699	1.22e-08	***
Crime vs control	0.025632	0.003557	7.206	5.94e-13	***
Fantasy vs control	0.028703	0.003386	8.478	<2e-16	***
Sci-Fi vs control	0.025641	0.003655	7.015	2.36e-12	***
Horror vs control	0.021062	0.003307	6.37	1.93e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In all six comparisons, the hypothesis was confirmed and incorrectly labelled extracts were read more slowly with more time spent per character, with the differences for all genre conditions being statistically significant.

Survey Analysis

All results from the post-eye-tracking survey were encoded and mean and standard devations calculated for participants' self-reported reading times given for each genre. The results are reported below in Table 21.

Table 21: Survey responses

Genre reading times:	1h-	2h	3h	4h	5h+	M	SD
Romance:	24	4	4	2	2	1.7	1.2
Crime Fiction:	27	7	1	1	0	1.3	0.7
Fantasy:	24	6	3	2	1	1.6	1.1
Science-Fiction:	27	5	1	1	1	1.4	0.9

Horror:	33	3	0	0	0	1.1	0.3

Genre ratings:	Str. Dislike	Dislike	Neutral	Like	Str. Like
Romance:	1	3	5	17	10
Crime Fiction:	1	2	8	13	12
Fantasy:	1	8	7	15	5
Science-Fiction:	5	11	6	11	3
Horror:	4	5	8	17	2

An interesting difference in results is between self-reported reading times for the given genres and the associated ratings. *Romance* received the most self-reported reading times and also leads in the "like" rating as well as being high for "strongly like" as expected. *Crime Fiction* received high scores in terms of participants liking it, but quite low reading time scores. On the other hand, *Horror* received extremely low self-reported reading times but surprisingly high ratings for participant enjoyment. Overall the pleasure reading reported by participants clustered around reading *Romance* and *Fantasy* stories, with some *Crime Fiction* and *Science-Fiction* and virtually no *Horror*. *Romance* and *Crime Fiction* also scored significantly higher on the enjoyment ratings than the other three genres, which are fairly equal.

In addition to these ratings, participants were asked to indicate whether they believed that the genres of text they were told during the experiment were accurate or not and given space to comment. The free-form responses were codified into a value of 1, 2, or 3. A one indicated that the participant considered all labels to be accurate and non-problematic. Two indicated that a participant thought that one or more but not all genre labels were inaccurate, and a three indicated a participant reporting that all genre labels were inaccurate in their opinion. The results are summarised in Table 22, for both the participants who read the version with correct labels and the version with incorrect labels.

Table 22: Participants' reports of perceived genre label accuracy

	Code 1	Code 2	Code 3
Incorrect labels:	12	3	3
Correct labels:	16	2	0

Surprisingly, participants' awareness of the genre labels in the manipulated version was very low, with only 6 out of 18 (33.3%) noticing that the genre labels were not accurate, and only half that (16.6%) noticing that indeed all labels had been incorrect. Two participants reading the control were also confused by some, and reported feeling that one or several of the correctly labelled extracts did not appear accurate to them.

In this section I have displayed the two primary analyses which were performed on the data, as well as their results. My primary hypothesis regarding reading speed increases for text samples labelled with incorrect genres was confirmed, while my predictions regarding readers' awareness of the problematic genre labels were interestingly proven incorrect.

In the next section I will discuss these results, their consequences for predictive model theory, and also discuss further the interesting results of the survey and significant comments made by participants regarding the genre labels.

5.5 Discussion

In this section I will delve further into the results presented above and discuss the results and their consequences for predictive model theory. I will discuss the primary prediction, which was successfully supported by the data, and analyse some of the commentary from participants to further delve into the reasons why the participants did not appear to generally be aware of the genre labels being inaccurate, or at the least did not report thinking so.

Given the clear outcome of the eye-tracking data, it can be concluded that giving readers a textual extract while also explicitly stating its genre has an impact on the reading speed. If the explicitly stated genre is incorrect, readers need longer to read the same extracts, and this is true across at least the ten texts from five well established literary genres used in this study. This result directly supports my theory of genre plausibility. The mention of a specific genre within each introductory text modulated global model expectations tied to the genre readers believed they were about to read. Instead of fulfilling those expectations, the genre plausibility assumed by the extracts fulfils others. This leads to the active predictive model being unable to integrate to the active global model predictions. While it is unclear from

the eye-tracking data alone whether participants chose to accept the stated but incorrect genre labels, or chose to interpret extracts as the genre they would more easily recognize them to be, the comments given by participants in the free form survey question can help to draw conclusions.

The free form answers were the largest surprise of the experiment, but also help to confirm that both immediate prior knowledge and well established schematic knowledge were used. Participants were unexpectedly inaccurate in deciding that genre labels were incorrect and showed a tendency toward rationalizing the incorrect labels as correct. As stated in Section 5.4, only 16.6%, 3 out of 18, of participants in the experimental condition correctly answered that all genre labels were wrong. For most of the participants in this condition, the reported answer was that all labels were accurate, despite the eye-tracking data showing that reading times were significantly longer. The conscious processing of some participants in particular centred on specific extracts while others appeared to have taken no issue at all. The most interesting examples are as follows, all taken from participants who read version 2 in which all genre labels were incorrect, grouped from response codes 1 to 3:

Code 1:

- “I think all the labels fit really well.”
- “Yes I felt they were accurate.”
- “Yes I thought that they were accurate.”
- “I think they were accurate.”
- “Accurate.”
- “I think they seemed accurate.”
- “Yes I felt that the genre labels for each extract were largely accurate.”
- “Yes, I feel the genre labels were accurate for the extracts.”
- “Yes, I think they were accurate, although I tend to associate romance with the two people involved & not just gossip.”

Code 2:

- “Most of them seemed accurate. One said the extract was romance, and it didn’t seem to fit that genre.”

- “For the most part the genre labels appeared accurate. As their[sic] was mostly evidence for these genres in the extracts I read. However, one of the ones labelled as a ‘Romance’ did not necessarily feel this way to me. This may however not be applicable to the whole novel, just the extract I read.”
- “Not always, one extract can’t tell the reader about the whole genre of the book, but some definitely didn’t seem to fit. For example the extract labelled Romance seemed far more like a horror.”
- “They were mainly accurate, although one extract labelled “Romance” that took place on a space craft seemed a little out of place.”
- “Most of them seemed very accurate. For example, one stated it was Crime fiction and it included a plot about detectives investigating a murder which was very accurate. But, one was stated as a romance but didn’t really seem like it – lacking any sort of love in any form. Nevertheless, all of the other genre labels did seem very accurate; the content fitting the genre.”
- “The one that said it was from a horror story, but the extract was about a girl who had been on a date.”

Code 3:

- “No, I realized a couple of extracts in to the experiment that in each description the genre stated didn’t seem to fit what followed. I don’t think any of the extracts were labelled the correct genre.”
- “The first extract seemed to fit into the romance genre but the label stated otherwise and described it as a science fiction. This happened with every extract and therefore all of the genre labels were incorrect, resulting in me not reading them properly towards the end of the experiment.”
- “I did not think that the genre labels were right in almost any case at all. Around the fourth extract was one labelled crime, it was definitely more SF from the amount I read.”

While in themselves these comments would indicate that there is indeed a kind of conscious choice to interpret genre, the ones who thought that all labels were wrong represent a small part of the sample group. Among those who commented on specific examples, most only considered one or two incorrect. *More importantly, despite the vast*

difference in conscious attitude towards the genre labels in the statements, the eye-tracking data still definitively indicated slower reading times for each extract across all participants, which suggests that the conscious awareness of incorrect labels did not appear to change the participants' reading strategy. Being able to discount the factor of awareness in this way allows me to draw the conclusion that there must indeed be an element of schematic genre knowledge which led to a contradiction as I hypothesized. I do not however believe that it was necessarily the same process which led to longer reading times in every case. Taking the participant's comments into account, there appear to be several strategies which affected their reading processes.

In order to process the extracts, participants needed to somehow integrate the information being given by the introduction screens with their prior knowledge. Thanks to the results of Zwaan (1994) we have other evidence that readers interpret text differently based on the preceding context and what they have been told about or initially read. The activation of immediate genre expectations would also be in line with the actual script theory of Schank and Abelson, who originally proposed that scripts could be activated by seeing or hearing key words which were associated with the sequence (1977). In this case readers did not only see the actual key word which is the name of the genre, but it was embedded in a phrase specifically communicating to them that the following text belonged to it. A fact often not mentioned in literary linguistics is that readers are conditioned to identify texts in this way, as virtually all forms of published works, physical or digital, come with an identifier such as this on the cover or splash page, on the back cover, in synopses, or in the case of the Kindle store the titles will often be embellished by a marketing slogan which also mentions the genre. Considering the definition of genre as a class of communicative events, this type of declarative would surely form part of a reader's personal genre schema. Whatever information the reader has about the genre at this point would be contained in their long term memory as part of the global model, with the skeletal schema as I discussed in 5.2 active to deliver top down predictions. The input which follows now needs to somehow be integrated into those existing predictions, or have errors solved through additional stages of errors and predictions, or taken at face value where there is not enough information within the global model or the text itself to draw conclusions about the plausibility of the text.

The participants reading the version with correct genre labels overwhelmingly appeared able to integrate the textual information into what they were expecting based on the genre

label. For those participants who read incorrect labels, other strategies had to be used. Here each participants' own repertoire of genre knowledge and experience became important. In all cases, participants need somewhat longer to read the extracts. For some, this appeared to be a clear cut case, in which all extracts seemed somewhat out of place. In reality, given the quite fuzzy definition of genre to begin with and the concept that genre schemata are prototype structures which only have family resemblance structures between them, there is already an inherent difficulty in properly recognizing and differentiating genres. Let us consider the *Horror* extract from *I have the Sight* (Wood 2016). In the experimental condition, it was mislabelled as a *Fantasy* story. Overall, participants in the survey did not self-report reading *Fantasy* as much as other genre types, often not at all, and Fantasy was not rated as highly as other genres. Horror received even fewer self-reported hours being read per week and the lowest enjoyment scores overall. We can assume from this that most participants' *Fantasy* and *Horror* schemata would have been far less fleshed out and nuanced than readers who particularly enjoy both genres. It is possible that the mention of demons and demonic possession as well as the fairly magical description of this extract may have been able to satisfy the *Fantasy* schema of a reader who is not familiar with the *Fantasy* genre, or that at least it would receive the benefit of the doubt, at a conscious level. It is not a perfect fit, as the eye-tracking data shows but does not appear to elicit any errors so grave that any participant felt the need to remember it or point it out. There is however a partial fit, except for one strong element of the extract, which is a Christian chant being performed by the characters. As such, it is possible that enough of the *Fantasy* schema some participants had was satisfied by the extract that only the one remaining element would elicit an error and need to be resolved by a new predictive model as in Figure 24 below.

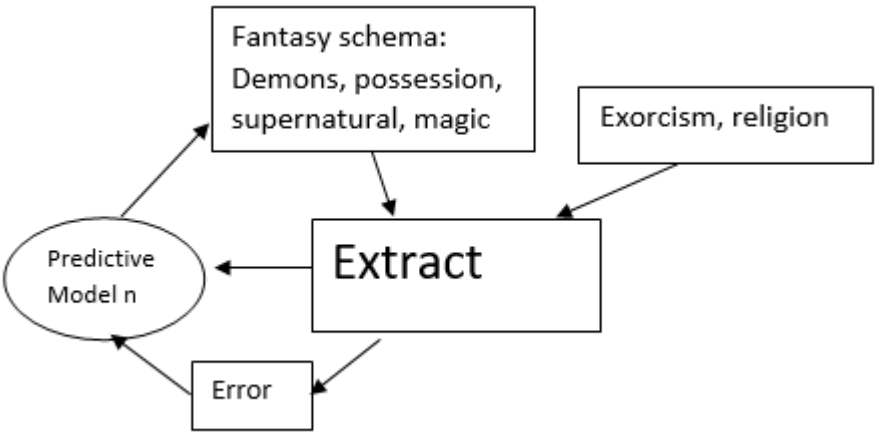

Figure 24: A skeletal Fantasy schema accounts for all but one thematic element of the extract

For some participants, perhaps there was a familiarity with some seminal exorcism horrors such as the obvious *The Exorcist* itself (Blatty 1971) or some of many complementary works produced since. This would have thrown doubt on the extract being a work of the *Fantasy* genre and required an additional step of prediction to deal with. Without additional information of commentary it is not entirely possible to give an answer what this step might be, but the two most likely answers would be that participants either took on the religious element as being an example of textual plausibility and formed a new predictive model which accepted that this particular *Fantasy* story also contained the somewhat unusual element, or that participants engaged in normalization.

In the first case, a new predictive model is formed which contains the accepted overlap in genre plausibility plus the new textual plausibility as in Figure 25 below.

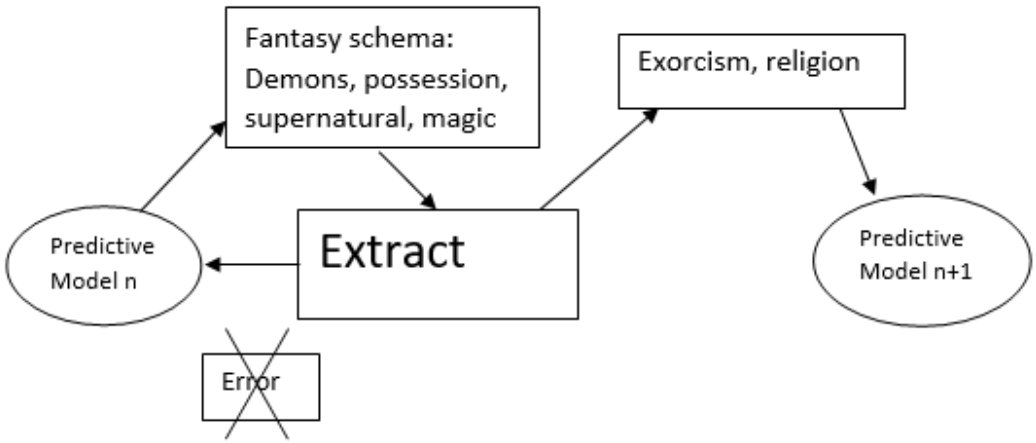

Figure 25: The unintegrated information becomes textual plausibility, forming new predictive model n+1

This strategy is only successful if a reader's existing knowledge of *Fantasy* texts does not contain any information which conflicts with the new textual plausibility. For some participants this may have been the case because they were so inexperienced with the genre that there simply was no information regarding the interplay between religious themes and *Fantasy*. For some others there may have been experience with many works of *Fantasy* including elements of fictional religions which shape the narrative to quite a degree, in which case the additional step may simply have been to include a real religion as potential trope. Operations like this are possible because it is already part of the prototype nature of genres that they contain innovation and additions of new textual plausibility. There is still the initial error however if a piece of information from the input does not match the initial expectations of genre plausibility. This is amplified the more an individual knows (or believes to know) about the genre in question and to what a degree this individual believes transgressions to be allowed. These are matters of individuality which are far beyond the scope of this thesis, but likely had an impact on the results of the experiment.

The second option mentioned is supported by some of the survey responses. I referred to it as normalisation because it is in essence the same strategy used in pragmatic normalisations, in which the reader "repairs" the errors arising from the input by changing the context around the input. This strategy is also based on the importance of initial textual information, and the communicative convention that publishers and book covers do not generally lie about the contents of their product. While texts can be deceptive through the

use of unreliable narrators, fictional publishers contained within the narrative and other tropes, this does not tend to extend all the way to the marketing stage of a book. When faced with a genre which a reader knows well and feels confident about being a member of the discourse community, it becomes more difficult to resolve errors in which this information conflicts with expectations of genre plausibility. In such a case, some participants may resort to taking the errors at face value as they are unable to integrate them into the global model and unwilling to accept them as new textual plausibility. The next best step is to assume that a mistake in the communicative purpose has been made. Said mistake would in this case be that the extract does not match the genre properly because the extract itself was either chosen poorly or does not contain enough information to confirm that it is that genre. This is a more conscious choice and is reflected in several comments given by participants, in particular with the genre they felt most comfortable passing judgement on: *Romance*. Participants pointed out *Romance* being used as an incorrect label more than any other genre used in the experiment, but notably in two very different ways.

The normalized versions were those in which participants wrote for example "[…]However, one of the ones labelled as a 'Romance' did not necessarily feel this way to me. This may however not be applicable to the whole novel, just the extract I read," and "Not always, one extract can't tell the reader about the whole genre of the book, but some definitely didn't seem to fit. For example the extract labelled romance seemed far more like a horror." Both of these comments suggest that the participants realized they could not integrate the extract itself into their global models, and were confident enough in their genre knowledge that they consciously were unable to accept any new textual plausibility to the contrary. They were however also unwilling to assume that the genre label itself was incorrect. The solution was to instead form a new predictive model that explains the discrepancy by assuming that the missing information which would make the text be the *Romance* they were told by the experiment it should be is in the story but simply not in the choice of extract. Essentially, they appear to have interpreted it as my error as the designer for picking an extract which did not correctly exemplify the genre it should be. This is likely based on the fact that participants rely on the genre they are told by a text as mentioned above, and of course as told by the experiment. They assumed that they would not be deceived by the experiment. It is entirely possible that due to this, their reports did not reflect their true beliefs that the labels were not correct. However if this conflict occurred during their reading, then it is entirely possible that this did influence and help cause the

slowdown in their reading speed which the data shows occurred. The discrepancy between expectation and initial textual information could be solved within a new predictive model in which the entire text contains whatever additional information their global model may require in order to accept it as a Romance as in Figure 26 below.

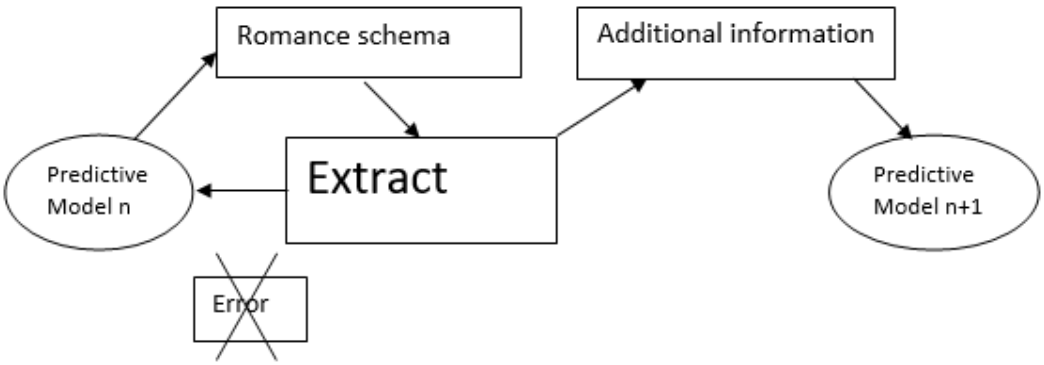

Figure 26: Error is cancelled out by assuming that additional information will be received, forming a new predictive model n+1

This approach would also constitute additional cognitive effort which in turn is expressed in slower reading speeds. It is an inferential approach which is made possible by the fuzzy and prototypical nature of genre in the first place, and the large amount of overlap between different genre types which may contain features that are central to some genres while only peripheral to others. *Romance* in particular is a genre whose central features appear in many other stories. Virtually all *Fantasy* and *Science-Fiction* stories, as well as many *General Fiction* and even *Crime* and *Horror* stories involve romantic sub-plots. It is a trope which has become a cliché in blockbuster movies that the hero always "gets the girl" at the end and oftentimes the romantic entanglement between a protagonist and other characters may even form or cause the actual climax of a narrative. For example, a protagonist's main quest within a *Fantasy* novel may be to free a captured loved one, or to win the heart of someone they are enamoured with. Even when such romantic tones only form minor sub-plots or are intended as characterization, they use many of the formal elements of what a pure *Romance* story is. This makes it possible to state that even though some chapters of a novel may for example fulfil the formal criteria of a *Romance* as a reader understands them, it could still overall be a different genre; almost any different genre. For the same reason, a normalization is possible since, as the participants stated, the rest of the novel chosen could have been the target genre, even if that was incorrect.

In order to identify these effects more clearly, future versions of the experiment could attempt to fully randomise the genre labels to see if some labels being truly correct influences the reading speed of the samples. In order to rule out the effects of participants normalising their reports due to believing they would not be deceived in an experiment it would be possible to inform all participants, in both control and experimental condition, that some genre labels may not be correct, and to observe how this influences both reading speeds and participant reports.

Regardless of the comments received, the eye-tracking data showed that participants read the incorrectly labelled texts statistically significantly more slowly, confirming my hypothesis that genre expectations based on being told a specific genre they were about to read influenced the reading process of the following texts. To my knowledge this was the first time such an experiment has been done with real texts, and I believe this shows great potential for future study.

In this chapter, I have discussed the results of the second eye-tracking experiment. I have concluded that the reading times measured directly supported my predictions and the predictive model theory of reading. They showed conclusively that readers faced with a specific genre label which does not fit to the text, reading times increased.

This, together with the results and discussion of chapter 4 completes our answer to the final research question: using predictive model theory, what does it mean to understand a fictional text describing events which never happened and how does this happen in a typical reading process? I believe these experimental results show that context is immensely important to our process of understanding, and that readers retain specific information about texts which are eventually entrenched in the global modal, and become predictions for future predictive models which are then compared to our global model expectations. Readers form complex knowledge structures of certain stories, which become genre concepts, and introducing mismatches between these and new texts creates new and unexpected errors which must be dealt with in some way. In my first experiment readers were freed from any suggestion of the genre, and thus were able to modify their expectations to include a sense of unreality and of a text which would not make sense. Here, I discovered the power of entrenchment of extended reading and genre structures, which did not allow participants to easily modify their global model expectations, and so

they expended further cognitive effort to suppress the errors generate by their genre predictions going unfulfilled.

In the next section, I will conclude this thesis by summarising all of the research I have presented so far and offering a final conclusion.

Chapter 6: The Contribution of Predictive Model Theory

6.1 Conclusions

In this final chapter I shall summarise my arguments, results, and give an overview of what my research has contributed to the field, and where it may go in the future.

As I have argued and shown in chapters 1 to 5, with the additional use of empirical eye-tracking data in chapters 4 and 5, there is good reason to adopt predictive model theory, which I have presented and argued for in this thesis, for textual understanding. The human brain follows the principle of Predictive Coding, which combines and synthesises top down processing in the form of predictions with bottom up input from our senses. It has great explanatory power for a number of previously unexplainable phenomena and as I have argued it also aptly explains many linguistic phenomena and principles of reading behaviour. The important addition when it comes to understanding texts and fictional stories is the idea of situation models. By taking the principles of predictive coding and the principles of situation model theory as well as many further cognitive and also linguistic theories I arrived at a particular framework, which I named predictive model theory. It states that our predictive brain sees the world as a series of events and actions with beginnings and ends and causal relationships, in which the primal causal relationship is that the existence of the world itself causes us to perceive it. Every closed sequence is a situation, resolved as one complete predictive model which an individual uses to internally resolve what is being perceived at a given moment, in a given context. Aspects of predictive models which are formed again and again are stored in our brain via long-term axonal connections between neurons. This knowledge forms our global model, our view of the world and how it works as based on what we have perceived and the neuronal activation patterns used to comprehend and interpret it. Fictional stories rely on the activation of this knowledge about the world, but also provide us with new knowledge. Creative choices of words describe new situations, ones which we are unable to perceive directly but which our brain may still resolve through a predictive model. The predictions built around such texts themselves become entrenched within our global model, since reading and telling stories are also part of our lives and of the world, in every sense. With this baseline, I believe I have provided a powerful baseline of theoretical principles with which to form predictions, and have gone on to test them in practice.

The principle of situational understanding as I have described it in this thesis is able to account for language comprehension in fictional texts by describing how readers activate knowledge and predictions to deal with novel fictional input as well as input based on the real world. Word meanings often rely on the context of the utterance as well as the words preceding and following and this too is resolved by predictive models which consider the sentence meaning and causal relationships between words and world, and words and sentence meaning. On the level of an entire narrative predictive models account for the present action within a text as well as for the actions in the context of the story as a whole, and in the case on genre plausibility for an entire cultural body of texts and the place of an individual story within it. At each level, it is the expectation of the individual, based on past experience, which defines those parts of any input which we perceive to be "errors", new information which does not fit into our existing worldview, as well as which strategy of cognitive processing will be able to resolve the error and form a new predictive model which can satisfy the global model and "explain" the problematic input. The context sensitive nature of my theory of situational understanding can account for this process in the abstract, while also explaining why different individuals will perceive different errors and accept different explanations within the same text in the same context. Each individual's unique reading experiences and knowledge surrounding the literature with which they engage will form a unique global model which responds differently to textual inputs. Due to culturally shared inputs many of the central prototypical linguistic devices in texts will overlap and be perceived similarly by individuals but the periphery and personal nuances in each case lead to a unique interpretation by each reader.

In practice, when encountering a text each reader begins by seeing the title and in the case of a novel the cover, activating certain expectations within their global model. At each stage of reading the knowledge about genre plausibility and texts in general shape the default predictions against which the read text is compared. As more input comes in it is categorised into minimal predictive models by utilising the one-to-one principle, forming action chains in which single agents perform single actions or change stats at singular spaces and points in time, and these minimal predictive models are refined into a predictive model which fully accounts for the characters, action, time and space of what is being read, while still conforming to global model expectations. Whenever new input does not match the current predictive model, error signals are sent along specific neuronal connections and a new prediction must be formed. Many of the errors elicited in natural reading are not of a

problematic kind, or anything which a reader does not understand, but changes of event structure, jumps between new and different settings and characters and all of the usual elements of narrative structure to which the predictive brain adapts easily. Sometimes, there will be an error within understanding, for which there are multiple coping strategies. In the case of a fictional explanation or a new type of trope or fictional situation, the reader may have to take on the new information as a form of textual plausibility. This is possible whenever a text gives a new piece of information or causal explanation which is new and cannot be integrated into the global model, but which does not conflict with other information available to the reader. When such textual plausibility is encountered and resolved many times, it will eventually become part of the reader's global model proper and form part of genre plausibility, a set of criteria which define things that are plausible when reading a particular group of stories and allow suspension of disbelief, but which may not be plausible in any other context. No matter what, the cycle of error and prediction continues until all error is explained, by whichever strategy.

Through the research in his thesis, I have built a theoretical tool which is able to explain text understanding both from top down to bottom up at any stage of this processing hierarchy by describing exactly how a real reader might be processing a particular piece of language at a given point, with their given prior knowledge. It is also able to predict how future text input may affect the same reader and why and how explanations of causal relationships and plausibility are interpreted the way they are, both when we are correct about the world and when we are demonstrably not. I have been able to show that classic hypotheses of textual difficulty with inconsistencies or contradictions leading to increased reading times are not correct in all cases, but that through the use of predictive models readers can adopt strategies to minimize reading times by adopting errors as expectations and forming new interpretations of difficulty as part and feature of a text. This can happen even when readers openly feel that portions of a text were confusing or difficult, but did not as a result refuse to read or go on, but ultimately acknowledge that the textual difficulty led to a shallower understanding despite a complete reading process, and this can be best explained by predictive model theory. When confronted with specific genres, I have been able to show that while readers slow down and are affected by incorrect genre labelling, this is rarely a conscious process, must happen at the online, predictive model processing stage. Here, I was able to show that reading times did slow, as textual extracts provided to them were still internally consistent and instead clashed with higher levels of expectations. Both my experiments were innovative designs, using a whole fictional text as

well as a test of genres within fictional literature with state of the art eye-tracking technology which to my awareness have not been done before and which I hope will help to further our understanding of the reading processes of fictional literature, and language in general.

References

Ayer, A. J. (1956). *The problem of knowledge / by A.J. Ayer.* London: Macmillan ; New York.

bank - definition of bank in English from the Oxford dictionary. (n.d.). Retrieved August 15, 2016, from http://www.oxforddictionaries.com/definition/english/bank

Banks, I. M. (1992 [1990]): *Use of Weapons*. London: Orbit.

Barsalou, L. W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes*, *18*(5–6), 513–562. https://doi.org/10.1080/01690960344000026

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(04), 577–660.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1281–1289. https://doi.org/10.1098/rstb.2008.0319

Busse, J.-P. (2004): Zur Analyse der Handlung, in: Wenzel, P. (ed.): *Einführung in die Erzähltextanalyse: Kategorien, Modelle, Probleme*. Trier: WVT Wissenschaftlicher Verlag trier (WVT-Handbücher zum literaturwissenschaftlichen Studium; Bd. 6).

Chisholm, R. M. (1957). *Perceiving: a philosophical study / by Roderick M. Chisholm.* Ithica, NY: Cornell University Press.

Clark, A. (2015). What "Extended Me" knows. *Synthese*, *192*(11), 3757–3775. https://doi.org/10.1007/s11229-015-0719-z

Clark, A. (Ed.). (2011). Précis of Supersizing the mind: embodiment, action, and cognitive extension (Oxford University Press, NY, 2008). *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *152*(3), 413–416.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204. https://doi.org/10.1017/S0140525X12000477

Croft, A. (2011). *Too Close For Comfort* (Knight & Culverhouse Book 1). Circlehouse. Kindle Edition

Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics [electronic resource] / William Croft and D. Alan Cruse.* New York: Cambridge University Press. Retrieved from http://site.ebrary.com/lib/uon/Doc?id=10110137

Currie, G. (1985). What Is Fiction? *The Journal of Aesthetics and Art Criticism*, *43*(4), 385–392. https://doi.org/10.2307/429900

de Vega, M., Glenberg, A. M., & Graesser, A. C. (eds.) (2008). *Symbols and embodiment: debates on meaning and cognition.* Oxford: Oxford University Press.

Dembski-Bowden, A. (2016). *Night Lords: The Omnibus* (Warhammer 40,000). The Black Library. Kindle Edition

Dijk, T. A. van, & Kintsch, W. (1983). *Strategies of discourse comprehension / Teun A. van Dijk, Walter Kintsch.* New York ; London: Academic Press.

Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and Surprise Determine Neural Population Responses in the Ventral Visual Stream. *The Journal of Neuroscience*, *30*(49), 16601–16608. https://doi.org/10.1523/JNEUROSCI.2770-10.2010

event - definition of event in English from the Oxford dictionary. (n.d.). Retrieved August 22, 2016, from http://www.oxforddictionaries.com/definition/english/event

Etwas – definition of etwas in German from the Duden. (n.d.). Retrieved February 6, 2021 from https://www.duden.de/rechtschreibung/Etwas

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203. https://doi.org/10.1016/S0010-0285(03)00005-7

Fillenbaum, S. (1974). Pragmatic normalization: Further results for some conjunctive and disjunctive sentences. *Journal of Experimental Psychology*, *102*(4), 574–578.

Frege, G. (1948). Sense and Reference. *The Philosophical Review*, *57*(3), 209–230. https://doi.org/10.2307/2181485

Friston, K. (2002). BEYOND PHRENOLOGY: What Can Neuroimaging Tell Us About Distributed Circuitry? *Annual Review of Neuroscience*, *25*(1), 221–250. https://doi.org/10.1146/annurev.neuro.25.112701.142846

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325–1352. https://doi.org/10.1016/j.neunet.2003.06.005

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. (2012). A Free Energy Principle for Biological Systems. *Entropy*, *14*(11), 2100–2121. https://doi.org/10.3390/e14112100

Friston, K. (2013). Active inference and free energy. *Behavioral and Brain Sciences*, *36*(03), 212–213. https://doi.org/10.1017/S0140525X12002142

Friston, K. J., & Stephan, K. E. (2007). Free-Energy and the Brain. *Synthese*, *159*(3), 417–458.

Friston, K., Thornton, C., & Clark, A. (2012). Free-Energy Minimization and the Dark-Room Problem. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00130

Garrod, S., Gambi, C., & Pickering, M. J. (2014). Prediction at all levels: forward model predictions can enhance comprehension. *Language, Cognition and Neuroscience*, *29*(1), 46–48. https://doi.org/10.1080/01690965.2013.852229

Gavins, J. (2007). *Text world theory [electronic resource]: an introduction / Joanna Gavins.* Edinburgh: Edinburgh University Press. Retrieved from http://site.ebrary.com/lib/uon/Doc?id=10435292

Gettier, E. L. (1963). Is Justified True Belief Knowledge? *Analysis*, *23*(6), 121–123. https://doi.org/10.2307/3326922

Goodkind, T. (1995 [1994]): *Wizard's First Rule*. New York: Tor.

Goodkind, T. (1996 [1995]): *Stone of Tears*. New York: Tor.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2013). *Halliday's Introduction to Functional Grammar* (4th ed.). Abingdon, Oxon: Taylor and Francis.

Hampe, B., Grady, J. E., & ebrary, I. (2005). *From perception to meaning [electronic resource]: image schemas in cognitive linguistics / edited by Beate Hampe in cooperation with Joseph E. Grady.* Berlin ; New York: Mouton de Gruyter. Retrieved from http://site.ebrary.com/lib/uon/Doc?id=10197215

Hebb, D. O. (1949). *The organization of behavior: a neuropsychological theory / D.O. Hebb.* New York ; London: Wiley.

Helmholtz, H. & Southall, J. P. C. (1962). *Helmholtz's treatise on physiological optics. Volume 3 / edited by James P.C. Southall.* New York: Dover.

Hobb, R. (2009). *Dragon Keeper (The Rain Wild Chronicles, Book 1).* HarperCollins Publishers. Kindle Edition

Hohwy, J. (2011). Predictive Coding and Binocular Rivalry. *I-Perception*, *2*(4), 340–340. https://doi.org/10.1068/ic340

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010

Holtzman, S. (2014). *Wittgenstein To Follow a Rule*. Hoboken: Taylor and Francis. Retrieved from http://Nottingham.eblib.com/patron/FullRecord.aspx?p=1702297

Hubbard, H. H. (ed.) (1991): *Aeroacoustics of Flight Vehicles: Theory and Practice. Volume 1: Noise Sources*. Hampton: NASA Langley Research Center (WRDC Technical Report 90-3052) (NASA Reference Publication 1258, Vol. 1)

Accessed at: Hubbard, H. H. (ed.): *Aeroacoustics of Flight Vehicles: Theory and Practice. Volume 1: Noise Sources*. NASA Technical Reports Server, n. d. Web. 19 May 2015 < http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19920001380.pdf>.

Ingarden, R. (1968): *Vom Erkennen des Literarischen Kunstwerks*. Tübingen: Max Niemeyer.

Iser, W. (1972): *Der Implizite Leser*. München: Wilhelm Fink.

Jack, B. N., & Hacker, G. (2014). Predictive Coding Explains Auditory and Tactile Influences on Vision during Binocular Rivalry. *The Journal of Neuroscience*, *34*(19), 6423–6424. https://doi.org/10.1523/JNEUROSCI.1040-14.2014

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, *29*(2), 133–159. https://doi.org/10.1016/0749-596X(90)90069-C.

Kripke, S. A. (1981). *Name und Notwendigkeit*. Frankfurt am Main: Suhrkamp Taschenbuch Verlag (surhkamp taschenbuch wissenschaft 1056)

Kripke, S. A. (2008). Frege's Theory of Sense and Reference: Some Exegetical Notes 1. *Theoria*, *74*(3), 181–218. https://doi.org/10.1111/j.1755-2567.2008.00018.x

Lakoff, G. & Johnson, M. (2003 [1980]): *Metaphors We Live By*. Chicago: University of Chicago Press.

Langacker, R. W. (1986). An Introduction to Cognitive Grammar. *Cognitive Science*, *10*(1), 1–40. https://doi.org/10.1207/s15516709cog1001_1

Langacker, R. W. (2009). A dynamic view of usage and language acquisition. *Cognitive Linguistics*, *20*(3), 627–640. https://doi.org/10.1515/COGL.2009.027

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567. https://doi.org/10.2307/2025310

Liasson, M. (2016). *Can't Stop Loving You*. Montlake Romance. Kindle Edition.

Longley, B. (2016). *What You Do to Me* (The Haneys Book 1) Montlake Romance. Kindle Edition.

Menzies, P., & Pettit, P. (1994). In Defence of Fictionalism about Possible Worlds. *Analysis*, *54*(1), 27–36. https://doi.org/10.2307/3328100

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567. https://doi.org/10.1016/j.jml.2012.01.001

Murakami, H. (2003): *The Elephant Vanishes*. London: Vintage

Murakami, H. (2011 [1993]): *The Elephant Vanishes*. (Kindle Edition) Vintage Digital

Naylor, M. B. (1986). A Note on David Lewis's Realism about Possible Worlds. *Analysis*, *46*(1), 28–29. https://doi.org/10.2307/3328741

Neurohr, B. (2019). A predictive coding approach to Text World Theory. In Neurohr, B. & Stewart-Shaw, L. (eds.) *Experiencing Fictional Worlds*. Amsterdam/Philadelphia: John Benjamins (Linguistic Approaches to Literature volume 32)

Nevill, A. (2011). *The Ritual*. Pan Macmillan. Kindle Edition

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111. https://doi.org/10.1162/jocn.2006.18.7.1098

object - definition of object in English from the Oxford dictionary. (n.d.). Retrieved August 22, 2016, from http://www.oxforddictionaries.com/definition/english/object

Park, H.-J., & Friston, K. (2013). Structural and Functional Brain Networks: From Connections to Cognition. *Science*, *342*(6158), 1238411. https://doi.org/10.1126/science.1238411

Plato (2001). *Theaetetus [electronic resource] / by Plato; translated by Benjamin Jowett.* Blacksburg, VA: Virginia Tech. Retrieved from http://site.ebrary.com/lib/uon/Doc?id=5000900

Pulvermüller, F. (2008). Grounding language in the brain. In de Vega, M., Glenberg, A. M., & Graesser, A. C. (eds.) (2008). *Symbols and embodiment: debates on meaning and cognition.* (pp. 85–116). Oxford: Oxford University Press.

Pulvermüller, F., & Preissl, H. (1991). A cell assembly model of language. *Network: Computation in Neural Systems*, *2*(4), 455–468. https://doi.org/10.1088/0954-898X_2_4_008

Pulvermüller, F., Shtyrov, Y., & Ilmoniemi, R. (2003). Spatiotemporal dynamics of neural language processing: an MEG study using minimum-norm current estimates. *NeuroImage*, *20*(2), 1020–1025. https://doi.org/10.1016/S1053-8119(03)00356-2

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*, *2*(1), 79–87. https://doi.org/10.1038/4580

Rescher, N. (1999). How Many Possible Worlds Are There? *Philosophy and Phenomenological Research*, *59*(2), 403–420. https://doi.org/10.2307/2653678

Robson, R. (2015). *London Large - Blood on the Streets* (London Large Hard-Boiled Crime Series). London Large Publishing. Kindle Edition

Ryan, M.-L. (1991). *Possible worlds, artificial intelligence, and narrative theory / Marie-Laure Ryan.* Bloomington: Indiana University Press.

Sanderson, B. (2014). *Words of Radiance: The Stormlight Archive Book Two*. London: Gollancz.

Sanderson, B. (2016). *Arcanum Unbounded: The Cosmere Collection*. Orion. Kindle Edition

Sanford, A. J. (2008). Defining embodiment in understanding. In *Symbols and embodiment: debates on meaning and cognition / edited by Manuel de Vega, Arthur M. Glenberg, Arthur C. Graesser.* (pp. 181–194). Oxford: Oxford University Press.

Sanford, A. J., & Emmott, C. (2012). *Mind, Brain and Narrative*. Cambridge: Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. Hillsdale, N.J.; New York: Psychology Press.

Searle, J. R. (1975). The Logical Status of Fictional Discourse. *New Literary History*, *6*(2), 319–332. https://doi.org/10.2307/468422

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(03), 417–424. https://doi.org/10.1017/S0140525X00005756

Spitzer, M. (2008). *Geist im Netz: Modelle für Lernen, Denken und Handeln* (1. Aufl. 2000. Korr. Nachdruck 2008 edition). Heidelberg u.a.: Spektrum Akademischer Verlag.

Squire, L. R. (2008). *Fundamental neuroscience [electronic resource] / edited by Larry Squire ... [et al.].* (3rd ed.). Burlington, Mass; London: Academic Press. Retrieved from http://www.myilibrary.com?id=254054

Stalnaker, R. C. (1976). Possible Worlds. *Noûs*, *10*(1), 65–75. https://doi.org/10.2307/2214477

Tchaikovsky, A. (2015). *Children of Time*. Pan Macmillan. Kindle Edition

thing - definition of thing in English from the Oxford dictionary. (n.d.). Retrieved August 22, 2016, from http://www.oxforddictionaries.com/definition/english/thing

Therriault, D. J., Rinck,M., & Zwaan, R. A. (2006). Assessing the influence of dimensional focus during situation model construction. *Memory & Cognition*, 34, 78–89.

Tomasello, M., & Kruger, A. C. (1992). Joint attention on actions: acquiring verbs in ostensive and non-ostensive contexts. *Journal of Child Language*, *19*(02), 311. https://doi.org/10.1017/S0305000900011430

Turner, M. (1998). *The Literary Mind: The Origins of Thought and Language* (New Ed edition). New York: Oxford University Press, USA.

Van Berkum, J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal Of Experimental Psychology-Learning Memory And Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

van Herten, M., Chwilla, D. J., & Kolk, H. H. J. (2006). When Heuristics Clash with Parsing Routines: ERP Evidence for Conflict Monitoring in Sentence Perception. *Journal of Cognitive Neuroscience*, *18*(7), 1181–1197. https://doi.org/10.1162/jocn.2006.18.7.1181

Vega, M. de, Glenberg, A. M., & Graesser, A. C. (2008). Framing the debate. In *Symbols and embodiment: debates on meaning and cognition / edited by Manuel de Vega, Arthur M. Glenberg, Arthur C. Graesser.* (pp. 1–10). Oxford: Oxford University Press.

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *Quarterly Journal of Experimental Psychology*, *31*(4), 591–597. https://doi.org/10.1080/14640747908400750

Wennekers, T., Garagnani, M., & Pulvermüller, F. (2006). Language models based on Hebbian cell assemblies. *Journal of Physiology-Paris*, *100*(1–3), 16–30. https://doi.org/10.1016/j.jphysparis.2006.09.007

Werth, P. (1999). *Text worlds: representing conceptual space in discourse / Paul Werth.* Harlow: Longman.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636. https://doi.org/10.3758/BF03196322

Wimsatt, W. K., & Beardsley, M. C. (1946). The Intentional Fallacy. *The Sewanee Review*, *54*(3), 468–488.

Wittgenstein, L. (2006 [1984]). *Werkausgabe Band 1: Tractatus logico-philosophicus / Tagebücher 1914-1916 / Philosophische Untersuchungen*. Frankfurt am Main: Suhrkamp Taschenbuch Verlag (suhrkamp taschenbuch wissenschaft 501)

Wood, R. (2016). *I Have the Sight (EDWARD KING Book 1)*. Rick Wood Publishing. Kindle Edition

Wu, L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, *132*(2), 173–189. https://doi.org/10.1016/j.actpsy.2009.02.002

Zwaan, R. A. (1994). Effect of Genre Expectations on Text Comprehension. *Journal Of Experimental Psychology-Learning Memory And Cognition*, *20*(4), 920-933.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). THE CONSTRUCTION OF SITUATION MODELS IN NARRATIVE COMPREHENSION: An Event-Indexing Model. *Psychological Science (0956-7976)*, *6*(5), 292–297.

Zwaan, R., Graesser, A., & Magliano, J. P. (1995). DIMENSIONS OF SITUATION MODEL CONSTRUCTION IN NARRATIVE COMPREHENSION. *Journal Of Experimental Psychology-Learning Memory And Cognition*, *21*(2), 386–397.

Zwaan, R. A., Radvansky, G. A., Hilliard, A. E., & Curiel, J. M. (1998). Constructing Multidimensional Situation Models During Reading. *Scientific Studies of Reading*, *2*(3), 199–220. https://doi.org/10.1207/s1532799xssr0203_2

Appendix A: Post-eye-tracking surveys

Experiment 1

Literature Exposure Questionnaire LC15 Participant Code:_____

Age:

Degree course: 1st year 2nd year 3rd year M.A. PhD

1. Please indicate your agreement by circling one option in the following:

I believe internal consistency is very important in stories.

Strongly disagree Disagree No opinion Agree Strongly agree

I prefer stories with a complex plot.

Strongly disagree Disagree No opinion Agree Strongly agree

I prefer stories with a simple plot.

Strongly disagree Disagree No opinion Agree Strongly agree

2. Please circle one response in the following:

How many hours per week do you spend reading for your work or education?

1 hour 2 hours 3 hours 4 hours 5 hours or more

How many hours per week do you spend reading for pleasure?

1 hour 2 hours 3 hours 4 hours 5 hours or more

What is your favourite genre of book or story?

General fiction Fantasy Science Fiction Mystery Thriller Crime Non-fiction

Other (please specify):_____

3. Please circle or give brief answers to the following, where applicable:

Have you read the story in the eye-tracking experiment before? Yes No

If yes, how long ago roughly?

_____
__

_____
__

Are you familiar with other stories by this author? If so, do you enjoy them?

_____
__

_____
__

Did you feel that there was anything strange about the way time worked in the story?

_____
_____
_____
_____
_____

Did you feel that there was anything strange about descriptions of space or location in the story?

_____
_____
_____
_____
_____

Did you feel that there was anything strange about descriptions of cause and effect within the story?

_____
_____
_____
_____
_____

Thank you very much for your time and participation!

By submitting this questionnaire I agree that my answers, which I have given voluntarily, can be used anonymously for research purposes.

Experiment 2

Literature Exposure Questionnaire LC29 **Participant Code:_____**

Age:

Degree course: 1st year 2nd year 3rd year M.A. PhD

1. **Please indicate your agreement by circling one option in the following:**

I believe internal consistency is very important in stories.

Strongly disagree Disagree No opinion Agree Strongly agree

I prefer stories with a complex plot.

Strongly disagree Disagree No opinion Agree Strongly agree

I prefer stories with a simple plot.

Strongly disagree Disagree No opinion Agree Strongly agree

2. **Please circle one response in the following:**

How many hours per week do you spend reading for your work or education?

1 hour 2 hours 3 hours 4 hours 5 hours or more

How many hours per week do you spend reading for pleasure?

1 hour 2 hours 3 hours 4 hours 5 hours or more

What is your favourite genre of book or story? (multiple answers possible)

General fiction Fantasy Science Fiction Mystery Thriller Crime Non-fiction

Other (please specify):_____

3. **Please circle as appropriate:**

On average, how many hours per week do you read stories from the following genres per week?

Romance

1 hour 2 hours 3 hours 4 hours 5 hours or more

Crime Fiction

1 hour 2 hours 3 hours 4 hours 5 hours or more

Fantasy

1 hour 2 hours 3 hours 4 hours 5 hours or more

Science Fiction

1 hour 2 hours 3 hours 4 hours 5 hours or more

Horror

1 hour 2 hours 3 hours 4 hours 5 hours or more

4. **Please indicate your preference by circling one option in the following:**

How much do you like reading stories from the following genres?

Romance

Strongly dislike Dislike Neutral Like Strongly like

Crime Fiction

Strongly dislike Dislike Neutral Like Strongly like

Fantasy

Strongly dislike Dislike Neutral Like Strongly like

Science Fiction

Strongly dislike Dislike Neutral Like Strongly like

Horror

Strongly dislike Dislike Neutral Like Strongly like

5. **Did you feel that the genre labels for the extracts were accurate? If not, can you think of an example that didn't seem right?**

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Thank you very much for your time and participation!

By submitting this questionnaire I agree that my answers, which I have given voluntarily, can be used anonymously for research purposes.