THE UNIVERSITY *of* EDINBURGH

Effects of context on semantic representations and mechanisms in humans and language models

Georgia-Ann Carter

THE UNIVERSITY *of* EDINBURGH

Doctor of Philosophy

Centre for Doctoral Training in Natural Language Processing

The University of Edinburgh

2024

Publication List

This thesis is in part based on the following articles that have been published during my doctoral studies.

Chapter 2 appears in: Carter, G. A., & Hoffman, P. (2024). Discourse coherence modulates use of predictive processing during sentence comprehension. *Cognition*, *242*, 105637.

Chapter 5 appears in: Carter, G. A., Keller, F., & Hoffman, P. (2024). Flexibility in conceptual combinations: A neural network model of gradable adjective modification. *PloS one*, *19*(7), e0307775.

If self-absorption, vague yearnings, and a nagging sense of incompleteness are sins, then surely I will burn for all eternity, and I will save you a seat.

~ Colson Whitehead

Abstract

The rapid and incremental nature of language processing is a central challenge for human cognition. Understanding how this challenge is met has resulted in a broad range of work focused on answering questions such as elucidating which mechanisms support processing and exploring which computational models serve as good approximates for language processing. Typically, this work has focused on semantic processing across single words or sentences. However, we now know that context plays an important role in how the semantic system processes linguistic inputs. In this thesis, I investigate the influence of context on semantic representations and mechanisms in humans and language models. As much of the literature has focused on single sentence contexts, I investigate context at both wider and narrower scales. The first branch of studies focus on this wider scale by investigating the impact of discourse coherence on predictive processing in both humans and Large Language Models (LLMs). The second branch of studies focus on the narrower scale by exploring the influence of context on pre-trained word embeddings in a perceptual property prediction task for both nouns and adjective–noun phrases. In addition, I investigated how a neural network encodes perceptual features in conceptual combinations. In the first branch of work, I found that human's lexical–semantic predictions are sensitive to discourse coherence, but especially so when semantic violations are present. From modelling, I found that LLMs are similarly sensitive to the relationship between context and a target sentence. This is in addition to coherence effects and their interaction with predictability, which suggests that the benefit of a highly coherent context extends beyond just lowering linguistic surprisal. In the second branch of work, I found reasonable performance for the perceptual prediction of the shape of a concept from word embeddings, but lower performance for the brightness of a concept. This was not impacted by contextual prompting for noun representations, though I did find a limited impact of context when predicting the brightness of adjective–noun pairs. This has implications for the interpretability of representations derived from language models and for debates on embodiment within human conceptual processing. In the final study, I found that neural networks can flexibly encode the modulation of conceptual features when nouns are modified with scalar adjectives. They do this by first learning to generate predictions based on the adjective, and then acquiring knowledge of how the adjective modulates particular nouns. In sum, this thesis adds greater depth to our understanding of how context influences language in humans and machines.

Lay Summary

Communicating what we mean is no small feat. Language allows us to speak our dreams into existence and share humorous anecdotes with friends. However, understanding how we process language so quickly and effectively is a huge challenge for researchers. Traditionally, much of how we think about how the meaning of language is represented in the brain is based on experiments that use single words or single sentences. However, we know that context can change our understanding of meaning—just think about a piano. When we are playing a piano, we may think more about the sound of the notes and the emotion that is conveyed to the listeners. Whereas if we must move a piano, we are much more likely to think about the size and the weight of the piano (to estimate how difficult this may be). What I am trying to convey here is that a word (or a concept) like "piano" can be associated with different features which makes up our knowledge of a piano. Depending on the context, we adjust how we understand meaning in the world. In this thesis, I present work that tries to answer how context influences the meaning of language. I then also ask whether computational language models are influenced by context in the same way. To do this, I define context across two different scales, one that is broader than a sentence and includes the wider discourse, and one that is narrower than a sentence and focuses on phrases.

The work presented in Chapters 2 and 3 is concerned with wider discourse context. In Chapter 2, I present three experiments where I have participants read three-sentence narratives. The first two sentences set up a scene, and the final sentence is either highly or less coherent with the overall narrative. In addition, within the final sentence, the target word is either predictable or not. This touches on the idea of why we are so quick to understand language when it is presented to us. One theory claims that it is because we are likely to predict what the next word will be and that we use other linguistic cues to determine how reliable these predictions will be. I am interested in the differences we see in how fast people read a word depending on whether it is predictable or not. When measuring the amount of time it takes people to read narratives like this, I found that people are sensitive to the coherence of a discourse, and will use this to determine how reliable the linguistic cue of predictability is later on in the sentence. In Chapter 3, I used this human reading time data and asked whether a computational model, GPT-2, would exhibit a similar pattern of results to humans given the changes in discourse context. Language models, like GPT-2, are commonly used now in everyday life to assist us. However, researchers also use them to answer questions about language without interference from other aspects of cognition, like vision or moral reasoning. This is how I use language models in this thesis—as a tool to tell us something about a system of language. In this work, I found that GPT-2 is sensitive to subtle topic shifts, similar to humans, and that it can even be used to predict human reading times. However, it does not fully account for the reading time differences I saw when coherence was manipulated. This suggests that humans engage in additional processing steps when reading coherent narratives.

My work in Chapters 4 and 5 focus on narrower scales of context. This is similar to the piano example I gave above, but this time I am interested in how the presence of

an adjective, like in the phrase “dark banana”, changes how we think about bananas (or any other object of your choosing). For example, when I think about a dark banana, I immediately associate it with the ripeness of the fruit, and so the adjective tells me something important about how I understand the piece of fruit being handed to me. In Chapter 4, I used word embeddings extracted from language models to ask whether they are influenced by context in the same way humans are by using them to predict perceptual feature ratings. I did this for both nouns and adjective–noun phrases. A word embedding is a representation of a word which computational models use. Here, I focused on perceptual features, such as shape and brightness, as this information is often taken in through the senses during our experiences in the world. This assumes that language models, with no ability to perceive the world, will have limited access to this information. Measuring the importance of a feature to a word is done through getting people to rate how typical a feature is, for example how typical “furry” is for a “cat”. I found good prediction for the prediction of shape, while brightness was less well represented, replicating previous results. I also found that adding a contextual prompt in the direction of the feature, such as “the shape of slippers”, had a limited impact on how well perceptual feature ratings were predicted. In Chapter 5, I explored how a neural network, another type of computational model, represents the change in perceptual features between a noun and an adjective–noun phrase. For example, how a model would reflect the change in brightness from a “banana” to a “dark banana”. Testing this could tell us something about how humans flexibly use perceptual information. I found that predictions made earlier on in training were first about the adjective, and then later on in training the influence of the noun appeared. I also found that the model is slowest to approximate appropriate brightness predictions for phrases where the adjective and noun have contradictory brightness associations, like “dark snow”. These findings provide further insight into how concepts may be combined and add to our understanding of concepts in general.

Building an insightful image of how we understand meaning over time is not easy. This thesis has attempted to answer a small part of this question by exploring how context influences how we understand language. Overall, I found that context is crucial for understanding in the moment, and that language models may not be as affected by context as humans are. In turn, this helps broaden our knowledge about the language system in the brain.

Acknowledgements

I want to express my sincere gratitude to and deep appreciation of all the people in my life that have helped me get to and through this chapter of my life. First, to my supervisors, Paul Hoffman and Frank Keller, whose insight, guidance and calm have made me a much more competent researcher. I also want to thank academic mentors from my past—Christina Kim and Mante Nieuwland—for giving me the space to try on new ideas. And my examiners—Itamar Kastner and Stefan Frank, for their insight and guidance on this work.

I would not have made it to this point without my family —Mum, Auntie, Nan and Grandad John. You gave me a space to be endlessly curious, in whichever direction I chose. And of course, I can't forget the many furry friends that have joined our family (Bonnie included).

This journey would also not have been possible without the friends around me. For those who have stuck around. Aileen Baird, from chemistry queens to PhDs, we've not done too badly. Poppy Palmer, whose kindness and giggles fill up a room. And Beth Oliver, whose quick wit catches me off-guard every time. To Victoria Poulton, for all of the laughter and 'za.

My time in Edinburgh has brought many lovely people into my life who have improved it dramatically. Henry Conklin, who has taught me that aesthetics are something worth pursuing. Carine Abraham and Austin Savill, who have provided a space to watch peak weebery and somewhere I feel at home. Irene Winther, whose excitement I find infectious and never fails to make me laugh. Special thanks also go to Tibbles for gracing me with the privilege of taking care of him. To Tom Hosking, for the many coffee chats and Nikita Moghe, for our regularly scheduled rants. There are many others (Anna, Aida, Atli and Eddie, to name a few) who have provided many evenings well spent. I never would have met these lovely people if it weren't for my CDT, now full of many inspiring people. I also want to extend my gratitude to Sally Galloway, for her kindness.

The members of office 3.50, who I regularly spent workdays and lunches with having engaging conversations about a wide range of topics, you made the days where it felt like no progress was made more bearable—Rohit Saxena, Ronald Cardenas and Jason Fong.

And finally, I would like to thank the neighbourhood cats of Edinburgh, of which I have befriended many, for bringing light into my life.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. The nature of semantic representations in humans and machines

As humans, we understand the meaning of novel sentences with astonishing capacity. The ability to build up a meaningful representation over time fundamentally relies on the semantic system, and the representations and mechanisms that support it. Semantic memory can be defined as the mental store of knowledge of the world, and is one of the cognitive foundations that enable us to understand and produce language (McNamara & Holbrook, 2003). Theories of semantic memory typically deal with how the meanings of words are mentally represented, and how these can be combined into more complex units. These theories should also account for the connection between word meaning and the world, with a clear account of the relationship between how semantic representations are formed and other cognitive systems, such as perception.

Traditionally, in the semantic memory literature, it has been theorised that our representations of concepts are context-free. For example, our understanding of "table" is an abstraction that is invariant across the many contexts in which we may encounter a table. One way that this has been represented is as a network, with concepts represented as nodes (Collins & Loftus, 1975; Collins & Quillian, 1969; Quillian, 1967). The number of links between concepts depends on the number of shared properties, with more links indicating a greater overlap in properties. The network separates knowledge about conceptual meaning from the linguistic representation and is organised by semantic similarity. The main retrieval mechanism is spreading activation, such that during processing, concepts are activated and this activation spreads from the node through the links in the network, with decaying activation over greater distances. An alternative traditional theory is the feature comparison theory. In this feature model, Smith and colleagues (1974) represent word meaning as a set of semantic features that vary on how important they are to a word's meaning. For example, the defining features of a "bird" include having a beak, having wings, and laying eggs. The authors used typicality ratings and association norms as measures of featural similarity between concepts, with a two-stage process required for verification (McNamara & Holbrook, 2003). However, a critical issue with these theories is that they assume that concepts are context-invariant.

In contrast, current accounts of conceptual knowledge recognise the importance that context plays in semantic comprehension. Such theories account for the degree of dynamicity that is inherent in our representations of concepts across multiple contexts (Barsalou, 1999). We see evidence of this from semantic priming studies (Connell & Lynott, 2014; Tabossi, 1988; Van Dantzig et al., 2008), where the way

in which a concept is processed depends on its preceding context. Yee and Thompson-Schill (2016) argue that context is a fundamental property that structures the semantic system. They argue that this means concepts are variant across individuals and that conceptual representations change over time within an individual. Traditionally, this was seen as a hindrance to theories of semantic comprehension as it is not immediately clear how communication occurs so seamlessly if we have variable representations of the objects and events that we discuss in language. However, the authors argue that is not necessarily as problematic as previously thought for semantic comprehension (Yee & Thompson-Schill, 2016). If this is the case, there is a current disconnect between much of the language processing literature, which has primarily gathered evidence on the semantic system in sentence-level context, and theories of semantic knowledge that recognise the influence of context.

Initial work by Rodd and colleagues to address this disconnect has demonstrated that context effects can last much longer than a sentence (Rodd et al., 2016). In an investigation on how previous encounters with ambiguous words may influence which meaning is most readily available upon processing, the authors found that just one encounter with an ambiguous word could bias the listener's interpretation of the word for up to 20 minutes (Rodd et al., 2013). These investigations highlight one possible way in which context influences the semantic system, demonstrating that further consideration of context is warranted and necessary for a full understanding of the cognitive mechanisms at play. The potential impact of contextual influences on the semantic system has wide-ranging implications for theories of semantic processing and the cognitive mechanisms that support it. It also has repercussions for the connection between language models and our understanding of semantic processing. Human processing data and knowledge of the semantic system are used as points of evaluation for language models. This is especially helpful for questions regarding the cognitive plausibility of language models. A better understanding of the semantic system in humans will enable more accurate evaluation of these models and their efficacy for testing hypotheses about semantic systems in general.

1.1.1. Scope of the thesis

This thesis addresses the question of how semantic representations in both humans and machines are influenced by contextual cues. Traditionally, context effects have been studied at the sentence level, but context can act at both wider and narrower scales than this. Here, I define wider scales to include effects at the level of discourse, and narrower scales to include phrasal level effects. This thesis includes two sets of investigations of contextual effects at these wide and narrow scales. The first set of investigations focuses on predictive processing. In **Chapter 2**, I present three self-paced reading experiments exploring how a discourse-level cue, coherence, influences the reading of predictable items. In **Chapter 3**, I explore whether surprisal from Large Language Models can explain the reading patterns observed from my behavioural experiments. In the second set of investigations, I focus on conceptual processing and explore the extent to which perceptual information can be represented and flexibly combined in language models. In **Chapter 4**, I ask whether additional context influences the representation of perceptual information in pre-trained word

embeddings, and the implications of this on theories of embodied cognition. In **Chapter 5**, I investigate how semantic flexibility in conceptual combinations can be computationally represented. In the current chapter, I will map out the relevant literature across four sections with introductions of the associated projects at the end of each section.

1.2. Predictive processing: experimental insights

It has been well established that language processing occurs very rapidly. For example, when hearing an utterance, comprehenders are capable of integrating each word into a representation of the unfolding sentence in an incremental manner (Marslen-Wilson, 1973; Swinney, 1979). Multiple studies have demonstrated a facilitation effect during language comprehension, such that sentence completions which are highly predictable given the previous context are processed more easily than those which are not. This has been demonstrated in the absence of an inhibition effect for anomalous or less predictable completions (K. I. Forster, 1981; Schwanenflugel & LaCount, 1988; Stanovich & West, 1981, 1983). Typically, the Cloze test is used as a measure of word predictability, where participants are given an incomplete sentence and asked to fill in the missing word (Taylor, 1953).

Common methodological approaches to studying this facilitation effect include self-paced reading, eye-tracking and neuroimaging, including electrophysiology. All of these measures are temporal in nature. Self-paced reading measures the amount of time it takes a participant to read text on a screen through keypress; this can include single words, or whole chunks of text (Jegerski, 2013). This method does not exhibit the temporal sensitivity that we see from more recent research methods, however, this type of data is commonly collected in combination with eye-tracking and neuroimaging methods and analysed as an additional behavioural measure. Eye-tracking records a participant's eye movements as they respond to stimuli presented on a screen. One particular paradigm often used to study language processing is the visual world paradigm—participants are presented with visual stimuli oriented in a grid-like manner around the screen and concurrently listen to spoken language. Here, the latency of eye movements to particular objects is measured, with the objects being specifically chosen to match or mismatch expectations formed from the auditory stimuli (Berends et al., 2016; Huettig et al., 2011). However, it has also been noted that findings from the visual world paradigm may conflate processes associated with the re-activation of semantic knowledge that was initially activated by the visual representation (Huettig, 2015; Nieuwland et al., 2018).

Electrophysiological measures are also a popular choice due to their temporal accuracy. Electroencephalography (EEG) measures electrical activity generated by the brain through electrodes placed on the scalp, while event-related potentials (ERPs) refer to the electrical activity that is elicited in response to a stimulus, such as language (Luck, 2012). ERPs are typically defined by their amplitude, latency and scalp topography (Kutas & Federmeier, 2011). The N400 effect is a negative-going wave between 200-600ms after stimulus onset and is maximal over centro-parietal electrode

sites. This effect is typically found in response to language with a semantic mismatch between the current input and previous context, as such, it is often the neural marker mentioned within the literature. However, many of the studies discussed report ERP effects at target word onset, while there have been arguments that to accurately show anticipatory processing effects, the ERP effect should emerge prior to target onset. One way to do this is to time-lock to a linguistically relevant cue that is found before the target word, for example, Wicha and colleagues (2003, 2004) took advantage of morphological gender agreement in Spanish, whereby the gender of the preceding article either matched or mismatched the gender of the expected target word. The trials were constructed so that no matter which article was present, there was a reasonable continuation. Therefore, if an effect is found between the conditions, it can more reliably be linked to disruption of an anticipatory process (Kutas et al., 2011). The authors found a separable effect of gender mismatch that emerged at the article, expressed as a widely distributed positivity (Wicha et al., 2004). Other studies using a similar paradigm have also demonstrated expectancy effects prior to target word onset (DeLong et al., 2005; Van Berkum et al., 2005).

1.2.1. Theoretical basis of facilitative effects

An important, and still ongoing, debate within the field has focused on whether this evidence of a facilitation effect is due to prediction or integration (Federmeier, 2007). Here, "prediction" can be defined as the mechanism that comprehenders use to activate linguistic information prior to processing input that carries that information (e.g., at word onset). It is this pre-activation which allows some processing to occur before word onset, and thus leads to a facilitative processing effect upon encountering the word (Pickering & Gambi, 2018). In contrast, "integration" can be defined as the mechanism that allows comprehenders to combine linguistic information that has been activated during processing of the input, with a representation of the prior input. As such, integration does not include pre-activation, and instead is primarily concerned with bottom-up processing. In this case, facilitation is thought to occur when a target word is highly compatible with the prior context and therefore easy to integrate. Another term that is widely used in the literature is "expectation". This can be interpreted as a broader umbrella term for when a comprehender anticipates semantic content, but may not have narrowed this down to a particular word (Van Petten & Luka, 2012). This demonstrates another distinction within the predictive processing literature – the linguistic level at which predictions can occur (Huettig, 2015).

Traditionally, researchers have argued against the notion that prediction is the central mechanism behind language processing, as most words that we encounter are not predictable and most of the contexts are not constraining enough for specific predictions to be made. This led to the theoretical assumption that bottom-up signals, such as phonological representations, were more important for language processing than any predictive mechanism prior to word onset (K. I. Forster, 1979; Marslen-Wilson, 1987). Other critiques of prediction include arguments regarding computational resources, noting that as prediction is only useful in a handful of circumstances (i.e., highly constraining contexts), it is a waste of computational resources that can be better applied to a different aspect of cognition (Jackendoff, 2003). Moreover, it has

been argued that a failed prediction is worse than none at all due to the additional processing required for resolution. A number of studies have reported an additional frontal, late positivity for mismatched trials, which has been theorised to reflect the additional difficulty faced by an actively disconfirmed prediction (Delong et al., 2011; Federmeier et al., 2007; Otten & Van Berkum, 2008; Thornhill & Van Petten, 2012).

However, many studies have demonstrated that people are faster at processing a more predictable word than a less predictable word. This has led to the argument that readers are rapidly comparing the meanings of upcoming words to the prior context but may not necessarily be predicting specific lexical items. Kutas and Hillyard (1984) demonstrated this with their related anomaly paradigm. In this paradigm, they contrasted high-cloze (i.e., predictable) congruent completions, anomalous completions and anomalous completions that are semantically related to the congruent word. They found that related anomalies elicited a larger N400 effect than the congruent endings, but smaller than the unrelated anomalies. From this, they argued for a featural semantic representation of words, where the sentence context facilitates processing for words that contain some number of features that match the preceding information (Kutas & Hillyard, 1984). This finding has since been replicated extensively across multiple linguistic features, such as grammatical gender (Fleur et al., 2020; Otten et al., 2007; Wicha et al., 2003, 2004) and definiteness (Burkhardt, 2006; Carter & Nieuwland, 2022; Fleur et al., 2020). A number of studies have investigated the further claim of form prediction, the suggestion that comprehenders not only anticipate meaning, but a specific word form, such as the visual or phonological representation (DeLong et al., 2005; Ito et al., 2016, 2018; Rommers et al., 2013). However, the evidence for form prediction is mixed, with suggestions that the mechanism is more limited in scope than meaning prediction (Ito et al., 2016, 2017; Nieuwland et al., 2018).

A range of theories have been proposed to explain the mechanisms behind this anticipatory processing effect. One such theory is that of spreading activation (Collins & Loftus, 1975; Quillian, 1969). Here, concepts are represented as nodes in a network, and the network is organized according to semantic similarity. The links between the nodes indicate the proportion of overlap in properties between the concepts. In essence, concepts are activated during language processing and activation spreads from one concept to another through these interconnected links. This effect suffers from decay, and so concepts which are further away (i.e., less semantically related) will receive less activation (Collins & Loftus, 1975; McNamara & Holbrook, 2003). This has been theorised to be a mostly automatic process. Other theories have utilised the idea of a dynamic network with spreading activation. Altmann and Mirković (2009) model language comprehension as a simple recurrent network (SRN) and critically discuss how this network learns linguistic structure over time. The authors argue for an equivalence between linguistic and non-linguistic representations, suggesting that it arises due to a shared substrate between language processing and event encoding (Altmann & Mirković, 2009; Huettig, 2015). Studies such as Rommers and colleagues (2013) lend support to this. This study investigated whether visual representations are also activated in an anticipatory manner. To do this, the visual world paradigm was used, and participants were either shown the target object, a shape competitor of the target object, or a control object. The authors found smaller negative ERP signatures in

the shape-related condition than in the unpredictable condition, suggesting that participants do activate perceptual features of a referent in an anticipatory manner (Rommers et al., 2013).

Another branch of theories have proposed involvement of the production system (Dell & Chang, 2014; Federmeier, 2007; Pickering & Gambi, 2018). Dell and Chang (2014) propose the P-chain framework, which is a connectionist framework that maps out interrelations between components of psycholinguistics. This was an attempt to connect language production, processing and acquisition in a framework where prediction occurs incrementally through engagement with the production system. A similar theory by Pickering and Gambi (2018), termed prediction-by-production, suggests that comprehenders initially determine the linguistic context of a speaker's utterance by activating production representations that correspond to those from the comprehension system during covert imitation. They then derive the underlying intention of the speaker's utterance by taking into account the non-linguistic context (e.g., shared world knowledge) and, finally, use their own production system to construct the underlying linguistic representations (Pickering & Gambi, 2018). The authors claim this to be a general mechanism that can be applied to predictions at all linguistic levels, and that it is largely an optional mechanism that supports comprehension but is not required for it occur. Federmeier (2007) also presents an account of prediction that is linked to production. Here, the author argues that neural asymmetries are more apparent in production, backed by evidence of hemispheric differences. Here, the idea is that left hemisphere processing is more expectation-based and thus biased towards predictive processing, while right hemisphere processing occurs in a more bottom-up manner (Wlotko & Federmeier, 2007). This account proposes that the left hemisphere bias arises from increased connectivity as both comprehension and production processes have been localised there (Federmeier, 2007).

While there is now a wealth of evidence for predictive processing during language comprehension, the majority of studies have demonstrated this using single-sentence materials. As such, little is known about how these local prediction effects are influenced by the broader discourse context in which we typically experience language. In **Chapter 2**, I address this by investigating the impact of a discourse-level cue, coherence, on local predictive processing effects. To this end, I ran three self-paced reading experiments that modulated the coherence between the target sentence and two-sentence preamble contexts, as well as manipulating the predictability of the critical word within the target sentence. Here, I found that comprehenders were sensitive to subtle shifts in discourse coherence, with downregulation of local predictive processing effects when the context is less coherent.

1.3. Surprisal Theory

Computational modelling of the types of facilitative effects described above has historically used measures from information theory. Hale (2001) proposed a quantification of the cognitive effort required to process a word in a given sentence

instantiated as the surprisal of the word in its given context. This notion of surprisal, which has been adopted from information theory where it is also known as Shannon information (Shannon, 1948), is defined as the log of the inverse probability of a word, *w*, appearing with a given context, *C* (see (1)).

$$surprisal = -log_2 P(w_i|C) \quad (1)$$

An overarching cognitive theory has been built based on surprisal, Surprisal Theory, which makes certain predictions about the patterns of human processing. In this way, surprisal has emerged as a metric of word-by-word cognitive load with the assumption that more probable structures are more difficult to disconfirm. Levy (2008) put forward surprisal as a measure of the re-ranking cost of incremental language processing. Here, the problem of incremental disambiguation when reading a given sentence is framed as an issue of allocating limited resources to possible analyses of the sentence. As such, the processing difficulty associated with a word, *w*, can be understood as the size of the shift in resource allocation during reading. One of the core predictions of Surprisal Theory posits lexical access as a causal bottleneck in the overall process of sentence comprehension. Another common metric that has been adopted from information theory is entropy (Shannon, 1948). Hale (2003, 2006) proposed the entropy reduction hypothesis which states that an incoming element is costly to process when it signals a change from a state of high uncertainty (e.g., multiple equiprobable predictions) to a state of low uncertainty (e.g., where a single prediction is most likely). Here, uncertainty (see (2) is quantified as the entropy of the distribution over the set of possible upcoming elements. These can be words, parses or other linguistic chunks; for the purposes of this thesis, I focus on words.

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x) \quad (2)$$

Entropy reduction (Hale, 2003, 2006) specifically focuses on the change in entropy between two words, $H_i - H_{i-1}$. If the entropy is reduced from one word to another, then communicative uncertainty has been reduced; whereas in cases where entropy increases, it is represented as zero, following the assumption that an increase in uncertainty should not impact processing (Hale, 2016). Theoretically, entropy reduction and surprisal are assumed to capture different aspects of information complexity, and therefore both can serve as metrics for quantifying predictability during sentence processing (Hale, 2016; Lowder et al., 2018). Moreover, surprisal theory is agnostic as to how $\mathrm{P}(\chi)$ is computed, the underlying model could use syntactic, semantic or both types of information.

1.3.1. Explanation of human processing data

Much of the literature using these metrics attempts to use them as an additional variable to explain human processing data (e.g., reading times (RTs) and ERP amplitudes) (Frank et al., 2013; Futrell & Levy, 2017; Levy, 2008, 2013; Staub, 2015; van Schijndel & Linzen, 2018; Zarcone et al., 2016). For example, Monsalve and colleagues (2012) compared two different model architectures, phrase structure grammars (PSGs) and recurrent neural networks (RNNs), in their estimation of lexicalised and unlexicalized surprisal. In this study, unlexicalized surprisal relates to the surprisal assigned to the part-of-speech (POS) tag. The authors found that lexicalised (i.e., word-based) surprisal is a significant predictor of RTs for naturalistic texts, outperforming unlexicalized surprisal. Surprisal has also been used to predict eye tracking data (Aurnhammer & Frank, 2018; Delogu et al., 2017; Demberg & Keller, 2008). For example, Aurnhammer and Frank (2018) evaluated the effectiveness of surprisal and entropy to predict human processing data compared to a novel metric, lookahead information gain. Lookahead information gain quantifies the information gained from processing a word, w_t, when probabilistically looking ahead to w_{t+1}. In this study, the authors compared SPR, eye-tracking and N400 data of naturalistic reading. All of the metrics were derived from Long Short-Term Memory (LSTM) recurrent neural networks (Hochreiter & Schmidhuber, 1997). A strong relationship between surprisal and N400 amplitudes and RTs was found.

This is a common finding in the literature, as many studies have now demonstrated a strong relationship between surprisal and N400 amplitudes, whereby words that elicit larger N400 amplitudes typically have higher surprisal values (Frank et al., 2013; Frank & Willems, 2017; Goodkind & Bicknell, 2018; Michaelov et al., 2021, 2021, 2023b; Michaelov & Bergen, 2022; Wilcox et al., 2020). Extending this to other ERP components, Frank and colleagues (2015) analysed six of them, including the N400, P600 and post-N400 positivity, collected from a naturalistic reading study. The authors compared surprisal and entropy measures extracted from different models (n-gram, RNN and PSG) in an effort to evaluate their cognitive plausibility. They found a strong relation between word surprisal and N400 amplitude, but not between any other component or information measure. They also found that estimates from n-grams and RNNs outperformed PSGs. Many of the studies discussed make use of naturalistic processing data where participants are passively reading a longer document of text, such as a book.

In contrast, Michaelov and Bergen (2020) investigated which experimental manipulations would elicit the same differences in N400 amplitudes and surprisal. Here, the authors found multiple linguistic phenomena that the RNN models could predict successfully. These included cloze, relatedness, semantic typicality and semantic anomalies. Some of the linguistic phenomena that were not well predicted included quantifiers and morphosyntactic anomalies. The authors conclude that this suggests some activation of semantic and lexical features that are indexed by the N400 cannot be entirely captured by exposure to linguistic input alone (Michaelov & Bergen, 2020). Recently, Wilcox and colleagues (2023) have addressed a notable gap in the literature by extending modelling with surprisal to more languages. The authors verified

existing predictions from surprisal theory across 11 different languages from five language families. These predictions include that surprisal and contextual entropy can predict RTs and that the relationship between surprisal and RTs is linear. From modelling multilingual eye-tracking data, the authors found strong crosslinguistic stability across the languages, confirming each prediction.

1.3.2. Relationship between surprisal and predictability

A pertinent debate within the surprisal literature questions the nature of the relationship between surprisal and predictability. Surprisal theory assumes a logarithmic relationship between surprisal and predictability, stemming from the idea that quick processing requires preparation (e.g., generating predictions about upcoming stimuli), but that preparation itself is expensive (N. J. Smith & Levy, 2008, 2013). Meanwhile, alternative theories from psycholinguistics assume a linear relationship. For example, one theory for prediction generation is a serial guessing mechanism, whereby if a prediction is confirmed and an expected word is processed, a fixed amount of facilitation is received and faster RTs are observed (Brothers & Kuperberg, 2021; N. J. Smith & Levy, 2013; Van Petten & Luka, 2012). Similarly, a linear effect of predictability could also be generated by parallel predictions; in this case, comprehenders would generate multiple predictions about the upcoming word, with preactivation of lexical features of a word assigned in proportion to their estimated likelihood. As such, greater levels of preactivation, and its subsequent confirmation, lead to greater levels of facilitation (Brothers & Kuperberg, 2021). In an initial exploration of this, Smith and Levy (2008) conducted a robust analysis, finding a reliable logarithmic effect of conditional word probability on RTs from both eye-tracking and SPR datasets. In a follow-up study, a greater number of linking functions were contrasted, however a logarithmic relationship was still found. Here, the authors concluded that their findings align with the idea that RT predictability effects are mediated by lexical predictability effects, supporting the notion of a causal bottleneck as proposed in surprisal theory (N. J. Smith & Levy, 2013).

Recent work has contested this logarithmic relationship. Brothers and Kuperberg (2021) conducted two behavioural experiments with different methodological paradigms (SPR and cross-modal picture naming) and ran a meta-analysis of previous eye-tracking studies to examine the linking function in depth. They found a linear relationship between lexical predictability and word processing times for all three experiments, which they take as support for a proportional pre-activation account. Meanwhile, Szewyczk and Federmeier (2022) examined the relationship between word predictability and contextual facilitation. Here, word predictability was indexed by either cloze probabilities collected from participants or GPT-2 surprisal, and contextual facilitation was indexed by the N400. The idea to supplement human-generated cloze probabilities with GPT-2 surprisal values is due to the fact that probability differences between unpredictable words are difficult to measure using a Cloze task (Szewczyk & Federmeier, 2022; Taylor, 1953). As an initial step, the authors verified that the cloze probabilities and GPT-2 surprisal values were significantly correlated. In testing the relationship between word predictability and contextual facilitation, the authors found a generally logarithmic relationship, however early in the N400 time window, a linear relationship was also observed. Theoretically, the authors attribute both parts to the

activation of new semantic information (Szewczyk & Federmeier, 2022). Thus, the precise relationship between word probability and facilitation remains an active topic of debate.

1.3.3. Cognitive plausibility of computational models

Another important question within the surprisal literature concerns the cognitive plausibility of the models used. Modelling cognitive mechanisms using artificial neural networks has a long history in the cognitive sciences (Elman, 1990; Rumelhart & McClelland, 1987). Some of the architectures discussed stem from a desire to represent processing in a psychologically faithful way, for example, recurrent neural networks (Elman, 1990). However, deep learning models adopted from NLP have typically favoured performance rather than cognitive plausibility (Frank et al., 2019). Evaluating the psychological faithfulness of these models has only recently begun (Frank et al., 2019; Lake & Murphy, 2023). For example, Merkx and Frank (2021) compared RNN and Transformer-based language models (LMs) on their ability to predict RTs and N400 amplitudes. The Transformer architecture consists of self-attention layers with a linear feed forward layer (Luong et al., 2015; Vaswani et al., 2017). The self-attention mechanism allows the model to directly "attend" to previous parts of the input, which is cognitively implausible (Frank et al., 2019). In the study, while ensuring equal LM quality for a fair comparison, the authors found that Transformers generally provided a better fit to the human data than Gated Recurrent Units, a type of RNN. This was attributed to attention-based computations which allows Transformer models to better fit self-paced reading (SPR) and EEG data (Vaswani et al., 2017). Further analysis demonstrated that the advantage is mainly due to better performance on longer sentences. The authors highlight that these results raise the question of how good recurrent models are as models of human sentence processing if they are outperformed by a cognitively implausible model.

In a similar vein, Wilcox and colleagues (2020) conducted a broad evaluation of modern computational models as predictors of human reading behaviour. The authors compared the models on both architecture and the amount of training data provided, with evaluation on eye-tracking and SPR datasets (Futrell et al., 2017; A. Kennedy et al., 2003). The authors found a generally linear relationship between word-level surprisal and human reading times, and that model architecture has a substantial influence on performance. Another recent analysis of model architectures focused on the finding that larger, pre-trained Transformer models perform worse at predicting human RTs than smaller Transformer-based models. Oh and Schuler (2023) conducted a detailed linguistic analysis of the relationship between LM perplexity and the predictive power of surprisal estimates. Perplexity is a common metric for evaluating the linguistic quality of language models (Oh et al., 2022). The authors suggest that the finding that larger pre-trained LMs generate surprisal estimates which are less able to predict human RTs is due to memorisation. They caution against adoption of the "larger is better" assumption from NLP, and state that surprisal estimates from smaller pre-trained LMs should be used for modelling effects of sentence processing (Oh & Schuler, 2023).

As this review shows, there exist outstanding questions as to the degree to which surprisal estimates from Large Language Models (LLMs) are useful metrics for human language processing, especially given their lack of cognitive plausibility. Much of the work that has attempted to address this has focused on evaluation at the sentence level. In **Chapter 3**, I test how a LLM accounts for discourse-level effects on human reading, exploring how surprisal estimates from GPT-2 (Radford et al., 2019) map onto the human RT data collected in Chapter 2. I found that surprisal estimates are influenced by subtle topic shifts; however, surprisal does not fully account for the pattern of results observed in the RT data. This suggests that the benefit of a highly coherent context is not merely explained by lower linguistic surprisal.

1.4. Conceptual processing and embodied cognition

Predictive processing accounts aim to explain how we construct representations of sentences and wider events. For narrower contextual effects (e.g., phrasal-level effects), we need to consider the representations of individual concepts and their associated semantic information. Concepts held in semantic long-term memory are important building blocks of human cognition (Kiefer & Pulvermüller, 2012). They constitute the meaning of objects, events and abstract ideas (Humphreys et al., 1999; Levelt et al., 1999). However, it is important to note that the relationship between a word and a conceptual representation is not 1:1, as exemplified by language change (Dove, 2022; Kiefer & Pulvermüller, 2012). For example, the word "wicked" used to have connotations of evil, whereas it can now be used to indicate enthusiasm. A core theoretical debate within the cognitive sciences relates to how words get their meaning, also known as the "grounding problem" (Harnad, 1990). It is exemplified in the thought experiment put forward by Searle (1980), known as the Chinese room, whereby a system, who can be thought of as a "person" who does not speak Chinese, answers questions in Chinese using a library and predefined rules. Externally, it appears as if the system "understands" Chinese, however no semantic understanding is present.

In turn, even the nature of conceptual representations themselves is still a matter of debate. The role of sensory and motor representations in defining concepts has been discussed controversially. Concepts have traditionally been viewed as abstract mental entities, separate from the perceptual and motor systems (Pylyshyn, 1980; Quillian, 1969). On these views, it is thought that the sensorimotor features of objects and events are transformed into a common amodal representation in the semantic system. However, more recent modality-specific approaches assume that concepts are essentially grounded in perception and action (Barsalou et al., 2003; Lakoff et al., 1999; Pulvermüller, 2005). This aligns with the research program of embodied cognition, which emphasises the role of the body in all cognitive processes (Glenberg, 2010; Shapiro & Spaulding, 2024). For example, a number of studies have found evidence of sensorimotor simulation during cognitive tasks (Glenberg & Kaschak, 2002; Kaschak et al., 2005; Kaschak & Glenberg, 2000; Stanfield & Zwaan, 2001). Stanfield and Zwaan (2001) found a processing advantage on an image recognition task when the objects in the image matched the orientation of the object in a previous sentence. Neuroimaging has provided further support for sensorimotor simulation by

showing that the processing of words recruits the same neural regions that are activated during processing of the referent (Kan et al., 2003; Kaschak et al., 2005; Martin & Chao, 2001; Pulvermüller, 1999).

A number of criticisms have been presented against the idea that embodiment is crucial for conceptual processing. One such criticism argues that even a clear definition of what embodiment and grounding refers to is lacking (Chatterjee, 2010; Hauk, 2016; Mahon & Caramazza, 2009; Meteyard et al., 2012; Wilson, 2002). Another issue pertains to the representation of abstract concepts (Galetzka, 2017; Hauk, 2016). If conceptual processing is grounded in sensorimotor experience, then how can we form representations of concepts that have no concrete form such as "justice"? One answer provided by supporters of embodied cognition suggests that metaphors provide a way to ground abstract concepts. For example, Lakoff and Johnson (1980, 2008) argue that metaphor usage mediates conceptual understanding via sensorimotor activation that reflects literal word meaning. Alternatively, others have claimed that representations of abstract concepts may be based on simulations of the situations in which they occur (Barsalou, 1999; Wilson-Mendenhall et al., 2011). Whether or not this is the case, it remains that abstract concepts without corresponding sensorimotor components are a challenge for embodied cognition accounts (Dove, 2016). In addition, there has also been debate about the extent to which sensorimotor activation is vital for language comprehension. For example, sensorimotor impairment leads to only mild semantic processing deficits. Clinical studies with patient groups have demonstrated similar accuracies between patients and controls on semantic judgment tasks, as well as intact object recognition abilities in patients with motor impairments due to Parkinson's disease and apraxia (Kemmerer et al., 2013; Mahon & Caramazza, 2005). A theoretical perspective that can reconcile the mixed results regarding embodied language processing is the Hub-and-Spoke model, a theory of semantic memory that suggests the semantic network constitutes both modality-specific cortical regions and an amodal hub that integrates this multimodal input (Lambon Ralph et al., 2001, 2007; Patterson & Lambon Ralph, 2016). Evidence from neuroimaging indicates the anterior temporal lobes (ATL) as a possible neural correlate for this type of hub (Lambon Ralph et al., 2010; Visser et al., 2010).

1.4.1. Concepts in combination

In these debates about conceptual processing, less attention has been given to conceptual combination, the process whereby a new concept is created from pre-existing concepts (Coutanche et al., 2019). This ability of constructing complex concepts from simpler ones is critical for many aspects of cognition, and it has been suggested that some processes which mediate the integration of simple concepts into more complex ones may also be involved in the integration of sensory features into simple concepts (Coutanche et al., 2019). One way to characterise conceptual combinations is by how they are understood. Attributive combinations are feature-based, for example a "canary crayon", which is understood by taking a property from the modifier, mapping this onto a dimension of the head and integrating them. In this example, we take the colour feature from "canary", a bright yellow bird, and apply it to "crayon", a type of writing utensil, to correctly understand the referent of a bright yellow

crayon (Coutanche et al., 2019; Murphy, 1988; Springer & Murphy, 1992). Alternatively, there are also relational combinations, where understanding the relation between the items is critical, for example a "crayon box" (i.e., a box where crayons are stored). Multiple different relations can exist between constituent concepts, and so the precise relationship between the constituents is important. In addition, it has been found that conceptual combinations of a particular relationship can prime other compounds represented in the same way (Estes, 2003; Estes & Jones, 2006; Gagné, 2001).

Historically, theories of conceptual combinations have framed them as the result of amodal operations in predicate-like structures (Fodor & Pylyshyn, 1988). Connectionist approaches inspired from this replaced predicate-like processes with statistical mechanisms (Coutanche et al., 2019; Pollack, 1990; Smolensky, 1990). Other alternatives based on embodied simulation have suggested that people combine multi-modal simulations of individual concepts into larger and more complex simulations (Barsalou, 1999; Wu & Barsalou, 2009). For example, Wu and Barsalou (2009) investigated how the perceptual properties listed in a generation task change after conceptual combination. Further insights from neuroimaging have suggested that a combined concept is represented across the same neural regions that represent its constituent concepts and their corresponding features. For example, it has been argued that semantic knowledge is instantiated as distributed patterns of semantic features across the neocortex, with evidence of distinct neural correlates for colour, shape and size (Konkle & Oliva, 2012; Martin, 2007; Tanaka, 1996; Zeki et al., 1991).

Meanwhile, from behavioural studies, it has been shown that the activation of conceptual features varies depending on the situational context (Barclay et al., 1974; Barsalou, 1982; Van Dam et al., 2010; van Dam et al., 2012; Yee & Thompson-Schill, 2016). It has been theorised that the activity levels of features contributing to a concept differ as a function of the weighting of conceptual features and of contextual constraints (Binder et al., 2016). As such, the process of integration between concepts is non-trivial as the same property can vary when combined with different concepts. For example, a colour term such as "red" can have different values when integrated with "face", "fire" or "truck" (Halff et al., 1976). Solomon and Thompson-Schill (2020) examine this type of conceptual flexibility using a three-pronged approach of behavioural, neuroimaging and computational investigations. The authors specifically focus on the feature of brightness and used adjective–noun phrases as a test case. Adjective–noun phrases are a valuable way to isolate the integration process of conceptual combination as they are independent of additional processes of property selection (Bemis & Pylkkanen, 2011). They asked participants to rate the brightness of unmodified nouns (e.g., "paint") and adjectival modified nouns (e.g., "dark/light paint"). The authors found that adjectives ("dark/light") had a larger effect on the perceived brightness of an object when there was a high level of uncertainty about the typical brightness of the object. For words with high feature uncertainty (e.g., "paint"), the adjective influenced the brightness ratings to a larger extent, whereas for words with low feature uncertainty (e.g., "snow"), it had less of an effect. They also found that a Bayesian model which incorporates feature uncertainty exhibited the best fit to the behavioural data, while neural correlates for this flexible feature combination included the left inferior frontal gyrus (LIFG) and the left anterior temporal lobe (LATL).

This research direction highlights the importance of work that considers the flexibility of combinatorial semantics. In **Chapter 5**, I introduce work that I have conducted which investigates how a neural network encodes the flexible nature of gradable adjectives in adjective–noun phrases, using the perceptual feature of brightness as a test case. I found that the flexible learning of gradable adjectives was possible, with predictions first based on the adjective alone and then modulated by the noun later in learning. I also found that the model outputs mimicked the type of non-additive feature modulation present in the human data. These results have implications for understanding how semantic composition may occur and help generate testable predictions for future work.

1.5. Distributional semantics and word embeddings

One of the key arguments for embodiment in semantics is that much of our understanding of words comes from sensorimotor processing of the experiences they refer to. It has been argued that these experiential aspects of word meanings cannot be inferred from language alone. One way to test this claim is to investigate the adequacy of semantic representations that are derived purely from language inputs. Computational linguistics and NLP provide useful examples of this type of representation in the form of embeddings.

Distributional semantics is one of the main approaches for modelling the lexical content of a word (Boleda, 2020; Lenci, 2008; Lenci et al., 2022). In this approach, words have similar meanings if they have similar patterns of usage and co-occurrence with other words, following the Distributional Hypothesis (Firth, 1957; Harris, 1954; Lenci et al., 2022; Sahlgren, 2008). These representations typically consist of high-dimensional vector representations that encode the lexical content associated with a given word. These are commonly referred to as word embeddings, and they can be extracted from a range of models, such as Distributional Semantic Models (DSMs) and LLMs. DSMs can be categorised into two types: count-based models and predict-based models (Lenci, 2018). Count-based DSMs use the co-occurrence frequencies between words in a given context to build their vector representations. For example, these contexts can be defined by documents, as exemplified in Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) and Topic models (Griffiths et al., 2007), a moving window as seen in the Hyperspace Analogue to Language (HAL) model (Lund & Burgess, 1996) or corpora as demonstrated by GloVe vectors (Pennington et al., 2014). The co-occurrence matrices formed can then be transformed into dense vectors using dimensionality-reduction techniques, such as Singular Value Decomposition (Landauer & Dumais, 1997). In contrast, predict-based models build vectors using shallow neural networks which predict the vector for a word, given its context. Here, context is usually defined as a window of words surrounding the target word. Popular examples of these include Word2Vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) and FastText (Bojanowski et al., 2017).

The word embeddings that can be extracted from DSMs are static in nature, meaning that the representation for a single lexical item is the same, regardless of context. This presents issues with words such as polysemes, for example “bank”, where an associated token can have multiple meanings. This is representative of a wider issue with static word embeddings when we consider semantic representations that are contextually variable, for example in conceptual combinations (French & Labiouse, 2002; Glenberg & Robertson, 2000). Initial uses of these word embeddings included estimation of semantic similarity or relatedness between words, however, they are more commonly used now to initialise representations for deep learning models. These types of pre-trained word embeddings help LMs to capture semantic similarities which are useful for downstream tasks, and highlights the importance of distributional semantics within NLP (Lenci et al., 2022).

1.5.1. Contextualised word embeddings and semantic transparency

Recently, contextualised word embeddings have gained popularity. For these embeddings, each word token in a specific context has a unique representation, overcoming a critical issue with distributional word embeddings. Contextualised word embeddings can be extracted from Transformer-based architectures, which take a set of isolated embeddings as input and produce a set of contextually-informed hidden embeddings (Vaswani et al., 2017). As such, they can encode relevant aspects of context using a self-attention mechanism, which facilitates direct interaction between each element in a set with all other elements (Lenci et al., 2022). These embeddings have achieved widespread popularity due to their ability to capture multiple linguistic features and context-dependent aspects of word meaning. However, there are concerns regarding the increased levels of complexity and opacity with regards to how the models themselves learn and what sorts of information contextualised word embeddings contain (Bender & Koller, 2020; Bisk et al., 2020; Lake & Murphy, 2023; Lenci et al., 2022; Merrill et al., 2021). For example, Bender and Koller (2020) argue that meaning cannot be learnt from form alone, and as such, text-only models are unable to learn meaning in principle. In this way, the authors argue that LMs lack reference and grounding in the real world, and thus suffer from the grounding problem, as discussed in Section 1.4. (Harnad, 1990).

In contrast, it has also been suggested that LLMs may have achieved some aspects of meaning, mirroring our understanding of meaning from cognitive science. Piantadosi and Hill (2022) argue that LLMs may have a notion of meaning derived from the way concepts relate to each other. They argue against the issue highlighted by Bender and Koller (2020) by acknowledging that not all terms have a distinguishable referent (Wittgenstein, 1953). The authors then link this with conceptual roles, emphasising that concepts are built in a fluid and context-dependent manner (Barsalou, 1983; Casasanto & Lupyan, 2015). Similar arguments have also been put forward by Lake and Murphy (2023) in their comparison of how humans and machines represent word meaning. Evidence that has been put forward to support this includes research using human processing data as a point of evaluation. For example, Grand and colleagues (2022) developed a semantic projection method that enables word embeddings to be represented on featural scales (e.g., size) and found that they can

recover similar human judgements of concepts when organised on these scales. Abdou and colleagues (2021) explored the relationship between colour concepts in a similar manner. Here, the authors extracted word embeddings from BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020) and mapped them onto a 3D colour space, CIELAB, using either Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) or a linear mapping. The authors found that the structures recovered from the word embeddings for colour terms were aligned with the structure of the CIELAB space, meaning that some aspect of colour perception was contained within the word embeddings extracted from text-only LMs (Abdou et al., 2021). These results demonstrate that some perception-based aspects of semantic knowledge are transferable without explicit grounding.

This examination of the internal semantic content of word embeddings has also been conducted with DSMs (Andrews et al., 2014; Pereira et al., 2016). For example, Riordan and Jones (2011) conducted a comparison of feature-based (i.e., uses human-generated features to encode conceptual information) and distributional models. The authors used a semantic clustering task, where the aim was to recover similarity structures. They found that semantic category information was redundantly coded in both perceptual and linguistic experiences, with the feature-based and distributional models achieving similar clustering patterns. In a similar vein, Louwerse and colleagues have conducted numerous studies on what sorts of information can be learnt from language statistics alone. From these studies, it has been found that distributional embeddings can contain information regarding the geographical location of cities (Louwerse & Zwaan, 2009), the vertical location of objects in the world (Hutchinson & Louwerse, 2013) and can encode a range of perceptual features (Louwerse & Connell, 2011). As a unifying theory, Louwerse (2007, 2011) put forward the symbol-interdependency hypothesis. This provides an integrative perspective of language and perception that relies on the interdependency between symbols and embodied experiences, which can be represented by statistical models. In sum, these findings align with a branch of thought where distributional models are more than just tools. Günther and colleagues (2019) argue that DSMs are not merely engineering tools as they are formulated as cognitive theories, rely on psychologically plausible assumptions and can account for behavioural data. However, there remains concern about the use of DSMs to inform our understanding of human language processing, similar to those expressed about LLMs above (Glenberg & Mehta, 2008; Perfetti, 1998; Rogers & Wolmetz, 2016; Sahlgren, 2006).

As such, research into how faithfully representations from DSMs and LLMs convey human semantics is important. Frameworks from psychology can help in the rigorous testing of this, for example by using knowledge about the semantic system in humans to decode what meaningful semantic information is contained in these vectors. In **Chapter 4**, I investigate how well DSM/LLM word embeddings predict human judgements about the perceptual features of words. In two experiments, I tested if I could improve performance on a task predicting the perceptual feature ratings of concepts from embeddings by using the ability of Transformer-based LLMs to represent word meaning in context. In the first experiment, I focused on noun representations, where I compared decontextualised and contextualised Word2Vec and BERT

embeddings for a large set of concepts when predicting the perceptual features of brightness and shape. From this, I found very good prediction for shape, and more modest prediction of brightness. However, the addition of context had no influence on performance. In a second experiment, I used adjective–noun phrases and again predicted the feature of brightness, with and without context. Here, I found good prediction of brightness, with the non-additive effect of adjective modulation found to be represented within the word embeddings. Moreover, context had a limited impact on performance.

To summarise, the four studies in my thesis investigate how context influences the processing of meaning in language. They do this at a range of scales (discourse: Chapters 2 and 3; phrasal: Chapters 4 and 5) as well as in humans (Chapters 2 and 3) and computational models (Chapters 3, 4 and 5). Together, this thesis aims to develop our understanding of the complex processes by which combinations of words give rise to representations of meaning.

Chapter 2

Discourse coherence modulates use of predictive processing during sentence comprehension

2.1. Abstract

Context has been shown to be vitally important for comprehension. Lexical processing is facilitated when words are highly predictable given their local sentence context, suggesting that people pre-activate likely upcoming words to aid comprehension. However, this facilitation is affected by knowledge about the global context in which comprehension takes place: people predict less when in an environment where expectations are frequently violated. The current study investigated whether discourse coherence is an additional cue that comprehenders use to modulate lexical prediction. In a series of online, self-paced reading experiments, participants read target sentences preceded by short contextual preambles. Local facilitation effects were manipulated through the cloze probability of a critical word within the target sentence and discourse coherence was manipulated by varying the degree to which the target sentence was consistent with the information presented in the preamble. In the first two experiments, target sentences were read more slowly when they occurred in less coherent discourses, but no local facilitation effects were observed. In the third experiment, we strengthened the predictability manipulation by using semantically anomalous critical words. In this experiment, predictable words were processed more quickly and anomalous words more slowly when they occurred in highly coherent discourse. Our results suggest that comprehenders are sensitive to shifts in the topic of discourse and that they downregulate predictive processing when they encounter incoherence in the discourse. This is consistent with recent theoretical accounts suggesting that comprehenders flexibly engage in predictive processing, pre-activating semantic and lexical information less when their expectations are less likely to be reliable.

2.2. Introduction

Local context has a rapid influence on lexical processing. For example, there is a wide and established literature on the reduction in N400 amplitudes for target words that are semantically related to previous linguistic information (see Kutas & Federmeier, 2011; Kutas & Hillyard, 1980, 1984; Van Petten & Luka, 2012). When words are predictable based on their local sentence context, they are processed more quickly, which suggests that the processing mechanisms involved are anticipatory (Schustack et al., 1987). While much research has focused on understanding the nature of this anticipatory process, the extent of its influence and the types of information that contribute to it

remain under debate (Huettig, 2015; Van Petten & Luka, 2012). A range of linguistic properties have been implicated in these contextual influences, including the semantic properties of the preceding linguistic input (Kutas & Hillyard, 1984), the grammatical gender of a referent (Fleur et al., 2020; Nieuwland et al., 2018; Wicha et al., 2003, 2004), and the definiteness of a referent (Burkhardt, 2006; Carter & Nieuwland, 2022; Fleur et al., 2020). As such, it appears that cues at various levels of the language system are used during comprehension to facilitate the processing of upcoming information (Morris, 2006; Rabovsky & McClelland, 2020; Willems & Peelen, 2021). In the present study, we investigate the effect of cues at the lexical level (Experiments 1 and 2) and the semantic level (Experiment 3).

However, for comprehension to occur, information from the current sentence has to be integrated with prior information from the surrounding discourse context (Myers & O'Brien, 1998). To understand the discourse, people form a mental/situation model of it (Kintsch, 1998; van Dijk, 2006; Zwaan, 2016; Zwaan & Radvansky, 1998). We predicted that the status of this discourse model influences the processing mechanisms comprehenders engage in when dealing with sentence-level constraints. As such, the present study investigates how the use of local within-sentence constraints interact with the properties of the broader discourse in which the sentence occurs.

2.2.1. Evidence for Discourse Facilitation Effects

Individual sentences are typically preceded by a broader discourse context, which is in itself a rich source of semantic and conceptual information. The presence of a global discourse context facilitates processing for globally related concepts (Albrecht & O'Brien, 1993; Berkum et al., 1999; Camblin et al., 2007), through the formation of a mental model of the characters and events involved (van Dijk, 2006). A number of models of this process have been proposed. Kintsch's construction-integration model places a proposition as the key component, suggesting that a mental representation of a text is a network of these propositions, linked by shared arguments (Kintsch, 1998). Zwaan and colleagues have proposed a situation model account, whereby an event-based mental representation of the current state of affairs is formed (Zwaan, 2016; Zwaan et al., 1995; Zwaan & Radvansky, 1998). They propose that comprehenders update these representations autonomously during integration of the unfolding linguistic input. Depending on the overlap of semantic features from the incoming input and the current situation model, the current model is either maintained (i.e., high featural overlap), or updated to take new information into account (i.e., low featural overlap). It has been suggested that this is an automatic and continuous pattern-matching process (Myers & O'Brien, 1998). In the present study, we are interested in how this process of maintaining a model of the entire discourse interacts with the use of constraints within a sentence to facilitate comprehension. Throughout the study, we define the local context as the sentence currently being read, whilst the global context refers to the higher-level mental model that is maintained across multiple sentences (Myers et al., 1994; Zwaan, 2016).

So how does the wider global discourse impact the anticipatory processing advantages that are observed within local sentences? There is mixed evidence on this issue (Ledoux et al., 2006). In one important study, Hess et al. (1995) attempted to distinguish between lexical explanations and discourse-based explanations for context effects using a naming task. Participants heard short passages, saw the final word of the passage (the target word) on-screen and were asked to name this. The researchers manipulated how well the target word related to the global context and the local sentence, separately. They found that, without the addition of the prior context, naming latencies were facilitated if associatively related words were present in the sentence (e.g., "The *English/computer science* major wrote the **poem**"). This confirmed the presence of local facilitation effects. Subsequent experiments added global discourse contexts that were either related or unrelated to the final sentence. Naming latencies in the global related conditions were consistently faster than those in the global unrelated conditions. However, local facilitation effects were no longer reliably observed. In other words, a local facilitation effect was evident when the sentences were presented in isolation but was not visible when a supportive global discourse context was added in subsequent experiments. The authors concluded that discourse context plays a critical role in lexical processing and can override the effects of local constraints (for related findings, see Albrecht & O'Brien, 1993). However, one potential issue in this study is that the contexts used for the global unrelated conditions were not always the same across the local manipulation, which may have led to a confound.

Other studies have provided different perspectives on the roles of global and local context in lexical facilitation. For example, Traxler et al. (2000) manipulated both the association and plausibility of a target word, given the context (e.g.., "The *lumberjack/young man* carried/chopped the **axe** early in the morning"). In this example, "axe" is the target word, which is plausible in its immediate local context when preceded by the verb "carried" and is associated with the wider context when preceded by "lumberjack". Reading time measures were generally influenced by the plausibility of the immediate context but not the wider context, suggesting that global context does not override local plausibility constraints. Conversely, another study investigating how comprehenders combine information from both the local sentence and global context found that discourse-level information did impact the processing of upcoming words (Schwanenflugel & White, 1991). These authors conducted three experiments which varied in task (lexical decision vs. naming) and the distance of global information to the local target sentence (far vs. near). They concluded that discourse information interacts with information from the local context, with both sources of constraint benefiting the processing of upcoming words. These studies indicate that there is inconsistency in the literature on the relationship between global and local contexts, and their subsequent influence on processing.

Further evidence of the impact of global context on semantic facilitation effects has been found with neuroimaging. Camblin et al. (2007) used both eye-tracking and ERPs to investigate how discourse-level representations affect within-sentence priming. The authors manipulated semantic association between a local prime and the sentence-final target word, as well as congruency between prior discourse context and the target word. Their first experiment demonstrated a reduction in N400 amplitudes for

congruent conditions and associated conditions, while the second experiment also found shorter regression-path reading times for congruent and associated conditions. These findings suggest that both local and global constraints influence lexical processing. In subsequent experiments, the authors disrupted the presence of a coherent discourse by either presenting target sentences with no prior context or with an incoherent context made up of unrelated sentences. Under these circumstances, local association effects were stronger than when sentences were presented within a coherent discourse. From this, they concluded that discourse congruence had a robust and rapid effect on both ERPs and eye-tracking, whereby the facilitation from local lexical associations was strongest when there was no coherent discourse information to support processing (Camblin et al., 2007).

The possibility of an interaction between local and global contextual cues has been further explored by Boudewyn, Long and Swaab (2015). These authors conducted an ERP study using auditory stimuli to investigate how cues from both global and local contexts contribute to two levels of prediction (semantic features vs. lexical form). The authors created two-sentence mini-stories, manipulating global predictability of the critical noun given the discourse context and local consistency of the critical noun compared with an associated feature word within the local sentence (e.g., "Frank was throwing a birthday party, and he had made the dessert from scratch. After everyone sang, he sliced up some sweet/healthy and tasty cake/veggies that looked delicious"). The authors found a graded N400 response, whereby N400 amplitudes were influenced by both local and global manipulations. Moreover, evidence of an interaction suggested that support from one level was able to compensate for ambiguity and inconsistency at the other.

2.2.2. Mechanisms Supporting Contextual Facilitation

The studies discussed above provide a mixed view of how global and local cues interact to facilitate processing. Why do the effects of the global discourse seem to be so variable? One possibility is that use of predictive processing itself is flexible and varies depending on its likely utility. In other words, people may use cues from the wider experimental context to determine how valid their expectations are likely to be, and thereby modulate how strongly to engage in predictive processing. A number of studies have shown that extralinguistic cues can affect the degree to which local predictability influences processing, supporting the idea that predictive processing is dynamic and flexible (Brothers et al., 2017, 2019; Hagoort et al., 2004; Hald et al., 2007). For example, in an ERP study, Brothers et al. (2019) manipulated speaker reliability. Participants heard sentences read by different speakers. Critical sentences were identical for both speakers. However, high-cloze filler sentences were used to manipulate the characteristics of each speaker, such that the reliable speaker usually produced predicted sentence completions, while the unreliable speaker frequently produced completions that violated the listener's predictions. Brothers et al. argued that a purely automatic prediction mechanism would not lead to differences in N400 amplitudes across the manipulation of speaker reliability, as the linguistic context remained the same. In contrast, a flexible account would predict that comprehenders would be sensitive to the build-up of knowledge about the speaker across the

experiment and subsequently use that to tailor their expectations. Supporting the latter account, they found an interaction between speaker reliability and lexical predictability on N400 amplitudes, such that larger effects of contextual constraint were observed when the speaker was reliable, compared to unreliable. This suggests that language comprehension recruits a type of dynamic flexible expectancy mechanism, which is sensitive to extralinguistic cues about the likely value of context-based predictions. When participants listened to a speaker whose speech regularly defied expectations, they became less likely to predict words in advance.

This type of flexibility was also investigated in an earlier study by Brothers et al. (2017), which focused on whether top-down goals and strategies influenced predictive processing across both sentences and wider discourses. The findings provide further evidence that extra-linguistic context modulates predictive processes. In their first experiment, larger N400 modulations were found when participants were explicitly instructed to predict the upcoming target word, rather than just passively comprehend the stimuli. Moreover, in their second experiment, the authors found a smaller effect of predictability on reading times when the overall experimental environment was less supportive of predictive processing. The experimental environment was manipulated by changing the proportion of filler sentences that ended in highly predictable or less predictable endings. The assumption was that if comprehenders dynamically modulate their use of predictive processing, the effect of predictability should be reduced when the experiment contained a greater proportion of less predictable sentences. This is exactly what was found. As such, this finding provides further evidence for the influence of extra-linguistic cues on the degree to which context facilitates processing. It shows that local facilitation effects are attenuated when participants do not expect the sentence they are reading to provide reliable cues to its completion. In the present study, we investigated whether a similar attenuation effect would occur when readers encounter a shift in the global discourse topic.

2.2.3. The Current Study

In sum, there is a range of evidence with mixed results as to the interaction of global and local contextual cues on discourse comprehension, both in behavioural and neuroimaging studies (Federmeier et al., 2007; Just et al., 1982; Just & Carpenter, 1980; McDonald & Shillcock, 2003; Traxler & Foss, 2000). One possible explanation for the variety of findings is that the use of predictive processing is flexible and strategic, such that comprehenders are sensitive to cues about the reliability of predictive processing and use this to regulate the degree to which they use local context to facilitate word recognition. In the present study, we investigated this by manipulating the coherence of discourse in three-sentence written passages (see Figure 2.1). Coherence simply refers to the degree to which discourse forms a series of well-connected statements that form a meaningful whole (Ellis et al., 2016). When discourse lacks coherence (i.e., when we encounter unexpected or contradictory information within the discourse), we need to reconfigure the active situation model to accommodate this (Kintsch, 1998; Kuperberg, 2021; Zwaan, 2016; Zwaan & Radvansky, 1998). A lack of coherence could also indicate greater uncertainty about the reliability of local processing cues. If this is the case, people may use a lack of coherence in discourse as a cue to temporarily reduce the

influence of lexical predictability effects on processing until a stable situation model has been re-established. In contrast, if context facilitates comprehension in an automatic fashion, local context effects should not be sensitive to disruption at the discourse level.

To this end, we conducted three online, self-paced reading studies where we manipulated both the predictability of the critical word within the target sentence (as a local constraint), and the coherence of the target sentence with the preceding discourse (as a global constraint). This enabled us to investigate whether local facilitation effects are suppressed when comprehenders process less coherent passages. We recognise that anticipation of information across multiple linguistic levels (syntactic, lexical, semantic, and possibly phonological) can contribute to facilitation for expected continuations in sentences (Heilbron et al., 2022; Ito et al., 2016; Kuperberg, 2021; Kuperberg et al., 2020). Accordingly, we manipulated whether stimuli violated expectations at the lexical level (Experiments 1 and 2) or at the semantic level (Experiment 3). Violation of expectations at the lexical and semantic levels have similar effects on markers of predictive processing like the N400, though effects of semantic violations are often observed later in processing (DeLong et al., 2014; Haeuser & Kray, 2022; Nieuwland et al., 2020; Quante et al., 2018).

2.3. Experiment 1

2.3.1. Methods

Methods and analysis strategies for this experiment were pre-registered at https://osf.io/h73qg/?view_only=5c6dd5967cec4b0481b1713d31057cbe. Where procedures deviated from the pre-registered plan, we have noted this in the text. Further, all data associated with the current study is available on OSF at https://osf.io/a8j4c/?view_only=26fee0260d364b0b8ee65ded332ed969.

Participants

Ethical approval for all studies was granted by the School of Philosophy, Psychology & Language Sciences Research Ethics Committee at the University of Edinburgh and informed consent was obtained for all participants. Participants were recruited either on the University of Edinburgh's SONA platform and the Testable Minds online participant pool. Participants recruited through SONA were first-year Psychology undergraduates who were native speakers of English and were reimbursed with course credits. Participants recruited through Testable Minds consisted of a range of backgrounds; however, we required that the participants were based in the UK, had an approval rating greater than 90%, were native speakers of English and had a maximum age of 40. (Note: Testable Minds defines the approval rating as "the ratio of the number of approved results of a participant to the total number of completed studies"). In this way, we tried to minimise the demographic differences between each participant group. The use of an online study actually allowed us to reach a more diverse participant pool than an undergraduate sample (Enochson & Culbertson, 2015). Previous evidence from both reaction time experiments and comparison reviews have demonstrated that online

experiments are a viable method for data collection, including for self-paced reading paradigms (Anwyl-Irvine et al., 2021; Enochson & Culbertson, 2015; Johnson et al., 2022). Participants from Testable Minds were reimbursed with $5.60.

Prior to data collection, we performed a power calculation to determine the approximate sample size required. Because we had no prior estimate of the effect size for the interaction between coherence and predictability, we utilised a conservative effect size estimate of Dz = .4, an alpha of .05 and power of .80. This returned an estimated sample size of 52 participants. We used a two-step process of data collection whereby we pre-registered our intention to initially collect 52 participants' worth of data, check to see if there was a significant interaction between coherence and predictability, and if not, to collect another batch of 52. If there was already a significant interaction present, then we would have stopped data collection (Pocock, 1977). We used this approach to ensure that our study would be well powered, given that we did not have strong a priori expectations of the effect size. For our pre-registered analyses, we adjusted our alpha from 0.05 to 0.029 to account for this two-stage approach (Pocock, 1977). We report the results from our pre-registered analyses on our initial batch of participants in the Appendix (see Figure 7.2).

In total, we collected data from 134 participants who indicated having English as their first language, with 80 from the pool of Psychology students and 54 from Testable Minds (mean age: 23, range: 18-42, SD = 6.77). Our final sample size consisted of 84 participants, as 50 participants were excluded for the reasons stated below (see Analysis – Pre-registered).

Design

Participants read passages consisting of a two-sentence preamble and a target sentence. The study used a 2 x 2 within-participants manipulation of passage coherence and critical word predictability. The predictability of a critical word in each target sentence was manipulated using cloze probabilities. The coherence of passages was manipulated by varying the degree to which the target sentence was consistent with the context established by the preamble.

Stimuli

The stimulus set was developed from a prior dataset of sentences that included cloze completion values (Peelle et al., 2020). We filtered the sentences with a cloze probability of greater than 0.75 and selected a batch of 200 candidate sentences which became our target sentences. To manipulate predictability, we swapped the high cloze completion word for a less predictable continuation. We ensured that the less predictable word had the same part-of-speech as the highly predictable completion (see Table 2.1). We calculated the cloze probabilities for these less predictable words using the original Peelle et al. (2020) dataset. Where none of the participants in Peelle et al. provided a particular completion, we assigned it a cloze probability of 0. We added three "spillover" words to the sentences, after the critical word, in order to account for the nature of self-paced reading time effects, which are typically extended over time (Jegerski, 2013).

To manipulate coherence, we created high and low coherent two-sentence context preambles. For each target sentence, we began by creating a highly coherent preamble which provided a strong and highly meaningful context in which to process the target sentence. We then generated a low coherence preamble for each target sentence. The low coherence preambles were designed to be in the same general semantic domain as the target sentence to ensure they could be interpreted as part of a single discourse and such that the target sentence would either depart from the topic established in the preamble or would contain new or contradictory information. The high and low coherence versions of the preambles were constructed with a similar syntactic structure to ensure the stimuli remained as naturalistic as possible, and to avoid a change in processing demands (e.g., processing differences between active vs. passive constructions; Olson & Filby, 1972). As such, we had sentences spread across four conditions, with an example trial presented in Table 2.1.

Table 2.1. Example sentences across four conditions in Experiments 1 and 2.

Condition	Context Sentences	Target Sentence
High Coherence, High Predictability	Crime was a problem. Residents were starting to feel unsafe.	The police arrested the local gang for selling **drugs** on the corner.
High Coherence, Low Predictability		The police arrested the local gang for selling **toys** on the corner.
Low Coherence, High Predictability	The city was a safe place. It had a low crime rate and high employment.	The police arrested the local gang for selling **drugs** on the corner.
Low Coherence, Low Predictability		The police arrested the local gang for selling **toys** on the corner.

Note. Critical word in target sentence presented in bold.

Norming

120 participants rated the coherence of the preambles with the target sentences on a 5-point Likert scale (1: Not Coherent to 5: Very Coherent). They were presented with each preamble plus target as a single paragraph and were advised that coherence refers to the degree to which the sentences in the paragraph are meaningfully connected with each other and make sense as a whole, and to rate the passages accordingly. Each participant only saw one version of each trial (HCHP, HCLP, LCHP or LCLP). Participants were recruited from both the University of Edinburgh SONA recruitment platform and Amazon Mechanical Turk (AMT). 62 participants were rejected for failing attentional checks, thus in total, 58 surveys were analysed.

From analysis of the ratings for the 200 normed sentences, the best 160 sentences were selected whose HC and LC coherence ratings showed the largest difference, while controlling for other relevant psycholinguistic properties reported in Table 2.2. As such, we had a final set of 160 stimuli, with four versions of each passage across our four conditions. Properties of the stimuli in each condition are shown in Table 2.2. We ensured that HC and LC preambles were matched for length (number of words; t= -1.11, p= 0.27) and that HP and LP critical words were matched for frequency (t= 0.98, p= 0.33) and letter bigram frequency (t= 1.03, p= 0.31). However, the length of the critical words did differ slightly between the HP and LP conditions (t= -2.29, p= 0.02). We accounted for this within our pre-registered linear mixed-effects models for Experiment 1 by including critical word length as a covariate. For all other analyses, critical word length was included in our initial linear mixed-effects model as a fixed effect. We derived word frequencies from SUBTLEX-UK (van Heuven et al., 2014) and letter bigram frequencies from MCWord (Medler & Binder, 2005). Differences in coherence ratings between conditions was assessed using a 2x2 within-items ANOVA. As expected, HC trials were rated as more coherent than LC trials (F= 988.08, p= $1.36*10^{-131}$). HP trials were also slightly more coherent than LP trials (F= 18.87, p= $1.63*10^{-5}$), presumably because the presence of a less predictable critical word affected raters' perceptions of the coherence of the whole passage. However, there was no significant interaction between factors (F= 3.31, p= 0.07). To ensure that coherence differences between HP and LP could not account for the experimental effects, we modelled coherence using trial-specific coherence values in our analyses (in exploratory analyses for Experiment 1 and pre-registered analyses for Experiments 2 and 3).

Table 2.2. Average psycholinguistic properties of stimuli for Experiments 1 and 2.

Property	**HC/HP**	**HC/LP**	**LC/HP**	**LC/LP**
Coherence	3.83 (0.46)	3.57 (0.60)	2.46 (0.52)	2.36 (0.48)
Preamble Length	13.54 (2.05)	13.54 (2.05)	13.78 (1.92)	13.78 (1.92)
Critical Word Length	4.35 (1.16)	4.56 (1.09)	4.35 (1.16)	4.56 (1.09)
Critical Word Zipf Frequency	4.53 (0.80)	4.45 (0.88)	4.53 (0.80)	4.45 (0.88)
Critical Word Letter Bigram Frequency	1794.57 (1536.06)	1639.93 (1208.17)	1794.57 (1536.06)	1639.93 (1208.17)
Cloze Value	86.23 (6.59)	0.24 (1.17)	86.23 (6.59)	0.24 (1.17)

Note. Standard deviations are shown in parentheses. HC = high coherence; HP = high predictability; LC = low coherence; LP = low predictability.

Procedure

The study was coded in jsPsych (de Leeuw, 2015). Four lists, counterbalancing the assignment of each sentence to the four conditions, were created, alongside four additional lists with the trial order reversed to avoid order effects. Thus, each

participant saw one of the four versions of each passage. The 160 trials were split into four blocks, with a break between each block, the length of which was controlled by the participants. On each trial, the two-sentence preamble was first displayed on-screen in full. After reading this, participants pressed the space bar to trigger the target sentence. The target sentence was shown one word at a time, with each word centred on the screen. Participants moved to the next word in the target sentence by pressing the spacebar. Intertrial intervals were marked with a cross shown on-screen for 750ms.

To encourage participants to attend to the materials, yes/no comprehension questions were presented immediately after 20% of the target sentences. These were spread throughout the experiment and were evenly distributed across conditions. Answers were balanced for requiring information from either the preambles or the target sentences. Feedback was provided after each question to encourage participants to pay more attention if needed. Sentences were presented in black, 24-point Open Sans font on a white background. Comprehension question feedback was presented in coloured font depending on the outcome (green for correct and red for incorrect).

Analysis

Pre-registered. Prior to data collection, we submitted a pre-registration of our intended analyses. We divided each trial into three regions of interest for analysis: 1) preambles (two-sentence contexts), 2) pre-critical word (from the beginning of the target sentence up to, but not including, the critical word) and 3) critical word plus spillover (see Figure 2.1). Our main interest was in the processing effects on the critical word. However, in self-paced reading paradigms, behavioural effects are typically extended over time beyond the critical word itself (Jegerski, 2013). Therefore, it is the third ROI (critical+spillover) which is most informative for our hypotheses. However, for completeness, we also included a supplementary analysis of reading times for the critical word alone, without the spillover region. Our dependent variable was the mean reading time of the words within the ROI. Each ROI underwent winsorisation of the raw RTs occurring outside the range of 2 SDs from the mean, calculated for each participant. We also performed log transformations on RTs for the target sentence ROIs.

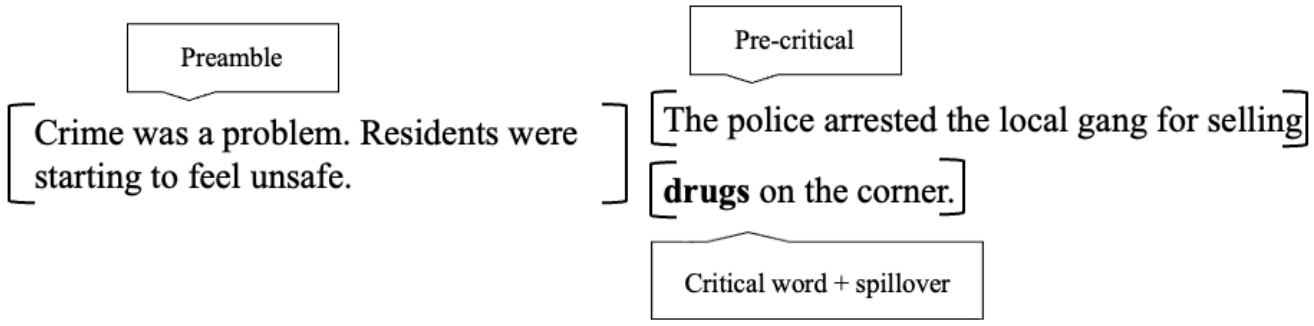

Figure 2.1. Example trial separated into ROIs used during analysis.

We fit linear mixed effects models in R for each ROI using the lme4 package (Bates et al., 2014; R Core Team, 2018). We included fixed effects of Coherence, Predictability and their interaction, as well as specifying length and Zipfian frequency of

the critical word as covariates in ROIs that included the critical word. For our categorical predictors, we used sum contrast coding. We also included by-participant and by-item random intercepts. For model fitting, we used a maximal approach, whereby if the maximal model did not converge, we altered the model in a step-wise fashion until convergence (Barr et al., 2013). These steps were, in order, removal of random correlations, removal of random slopes for the interaction terms and removal of random slopes for the main effects. Summary statistics were computed using lmerTest (Kuznetsova et al., 2017), which provides p-values using Satterthwaite's degrees of freedom method.

We also pre-registered exclusion criteria, which consisted of excluding a participant if their comprehension question accuracy was lower than 80% (n=50). Therefore, the final sample consisted of 84 participants. Additionally, we excluded trials that prompted an incorrect answer to a comprehension question (n=327).

Exploratory Analyses. After collecting the data and performing the pre-registered analyses, we re-analysed the data with an improved model, as described in the Results. As these changes were not pre-registered, we label these as exploratory for Experiment 1.

2.3.2. Results

Pre-registered Analyses

Following our pre-registered analysis plan, we ran four linear mixed effects models, each assessing the effects of coherence of the target sentence with the preamble, predictability of the critical word and their interaction. Estimated means for each ROI are presented in Figure 2.1.

Preambles. Participants had similar reading times for high and less coherent preamble contexts (β= -0.001, SE= 0.009, *t*= -0.14, *p*= 0.89), as well as for high and less predictable critical words (β= 0.004, SE= 0.006, *t*= 0.66, *p*= 0.51), with no interaction (β= -0.002, SE= 0.006, *t*= -0.29, *p*= 0.77). These results confirm that participants took a similar time to read the preambles across all conditions.

Pre-critical Word. For the pre-critical ROI, we found that processing was facilitated (i.e., faster reading times) when preceded by highly coherent preambles (β= -0.006, SE= 0.002, *t*= -3.07, *p*= 0.002). In contrast, there was no difference in reading times between high and low predictable conditions (β= 0.0007, SE= 0.002, *t*= 0.40, *p*= 0.69), which was expected as this region precedes the critical word. Further, no interaction effect was observed (β= -0.002, SE= 0.002, *t*= -1.10, *p*= 0.27).

Critical Word and Spillover. At the critical word and spillover region, our main ROI, there was no processing advantage for critical words preceded by high coherent conditions, compared to less coherent (β= -0.004, SE= 0.002, *t*= -1.56, *p*= 0.12), nor for high predictable critical words compared to less predictable (β= -0.005, SE= 0.003, *t*= -1.62, *p*= 0.11). Further, we found no interaction effect (β= 0.001, SE= 0.002, *t*= 0.54, *p*=

0.59). We did find marginal effects of critical word length (β= 0.007, SE= 0.004, *t*= 1.94, *p*= 0.05) and word frequency (β= -0.006, SE= 0.003, *t*= -1.80, *p*= 0.07).

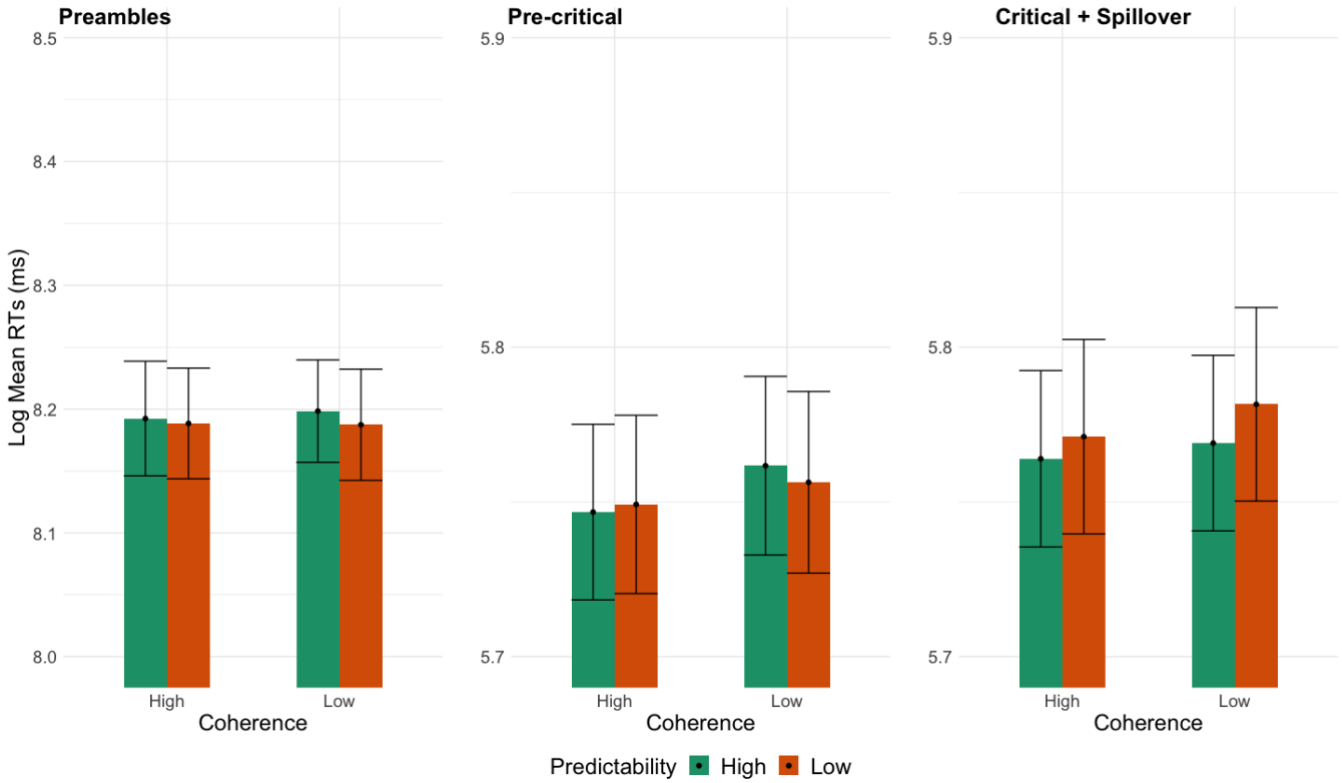

Figure 2.2. Experiment 1 pre-registered analysis results; error bars indicate the 95% confidence intervals.

Critical Word. At the critical word, there was no processing advantage for words presented with high coherent preambles, compared to less coherent preambles (β= -0.003, SE= 0.002, *t*= -1.47, *p*= 0.14). There was no effect of predictability (β= -0.005, SE= 0.003, *t*= -1.61, *p*= 0.11), nor was there an interaction (β= 0.002, SE= 0.003, *t*= 0.65, *p*= 0.52). We did find marginal effects of word length (β= 0.007, SE= 0.004, *t*= 1.81, *p*= 0.07), and frequency (β= -0.007, SE= 0.003, *t*= -1.96, *p*= 0.05).

In sum, for our pre-registered analyses for Experiment 1, we found that coherence facilitated processing at the pre-critical ROI, such that faster reading times were observed for high coherent trials compared to less coherent trials. However, this effect did not extend into our main ROI, the critical word and spillover region.

Exploratory Analyses

After completing our pre-registered analyses, we made a number of changes to our analysis pipeline to (a) improve our data exclusion criteria, (b) include more rigorous data pre-processing steps (e.g., residual reading times) and (c) treat coherence as a continuous variable. The modified analysis protocol used here was then pre-registered for Experiments 2 and 3.

Upon inspection of the data, we found that a small number of participants had very short, long or variable reading times for the preambles, suggesting that they were not paying sufficient attention to them. Hence, we added an additional exclusion criterion associated with preamble RTs. If mean RTs for the preambles was either less than 2s or more than 15s (n=11), or if the standard deviation of a participant's preamble RTs was more than 15s (n=6), we excluded that participant from the analyses. Finally, we removed all trials whose comprehension questions had a low mean accuracy across all participants (less than 70%), indicating that there was a general difficulty in understanding the sentence (n=7).

To account for effects of word length and individual differences in reading speeds, we fit an initial linear model on the log RTs from all words and ROIs, with a fixed effect of word length and random intercepts by participant. We then used the residuals from this linear model as the dependent variable in our main analysis models (Enochson & Culbertson, 2015; Ferreira & Clifton, 1986). The use of residual reading times to account for differences in word length and differences in individual participant's reading speeds has become widely accepted for self-paced reading paradigms (Lorch & Myers, 1990; Trueswell et al., 1994). For ROIs that consisted of more than one word, we took the average residual log RT per trial.

Further, we transformed our variable of interest, Coherence, from a two-level factor to a continuous predictor. We took the coherence ratings acquired during the prior norming experiment and scaled these. This gave a finer-grained representation of coherence, rather than binning it into two arbitrary groups. It also allowed us to fully disaggregate the effect of predictability from that of coherence (as HP trials had slightly higher coherence values than LP trials).

Finally, we conducted additional analyses including possible confounds as fixed effects. First, we included experiment half (first vs. second) and its interactions with predictability and coherence, to investigate whether participants' behaviour changed during the course of the experiment. Second, we included a binary fixed effect coding whether or not the critical word was also present in the preamble for a trial. Word repetition occurred for only a handful of the 800 variations of our stimuli; however previous research has demonstrated robust repetition priming effects on RTs (K. Forster I. & Davis, 1984). The inclusion of these additional covariates had no effect on the main results of any model; thus, we report the results of the models without the additional factors below.

Preambles. Estimated means for the new analyses are shown in Figure 2.2. For the preambles ROI, as expected, we found no advantage in reading times for highly coherent preambles (β= 0.002, SE= 0.01, *t*= 0.22, *p*= 0.82) or for trials with highly predictable critical words (β= 0.003, SE= 0.005, *t*= 0.60, *p*= 0.55), nor an interaction between the two (β= 0.006, SE= 0.009, *t*= 0.64, *p*= 0.52).

Pre-critical Word. For the pre-critical ROI, participants read more coherent trials significantly faster than less coherent trials (β= -0.009, SE= 0.003, *t*= -2.98, *p*= 0.003). As expected, there was no advantage for trials with highly predictable critical words (β=

0.001, SE= 0.001, *t*= 0.86, *p*= 0.39) and no interaction between coherence and predictability (β= -0.0002, SE= 0.003, *t*= -0.07, *p*= 0.94).

Critical Word and Spillover. For this ROI, we found that participants' reading times were affected by coherence, such that more coherent preambles facilitated language processing (β= -0.01, SE= 0.004, *t*= -2.47, *p*= 0.01). However, we found no difference in reading times between high and less predictable critical words (β= -0.003, SE= 0.002, *t*= -1.41, *p*= 0.16), nor was there an interaction between the two (β= 0.005, SE= 0.004, *t*= 1.29, *p*= 0.20). Further, there was no effect of critical word frequency on reading times (β= 0.0001, SE= 0.003, *t*= 0.03, *p*= 0.97).

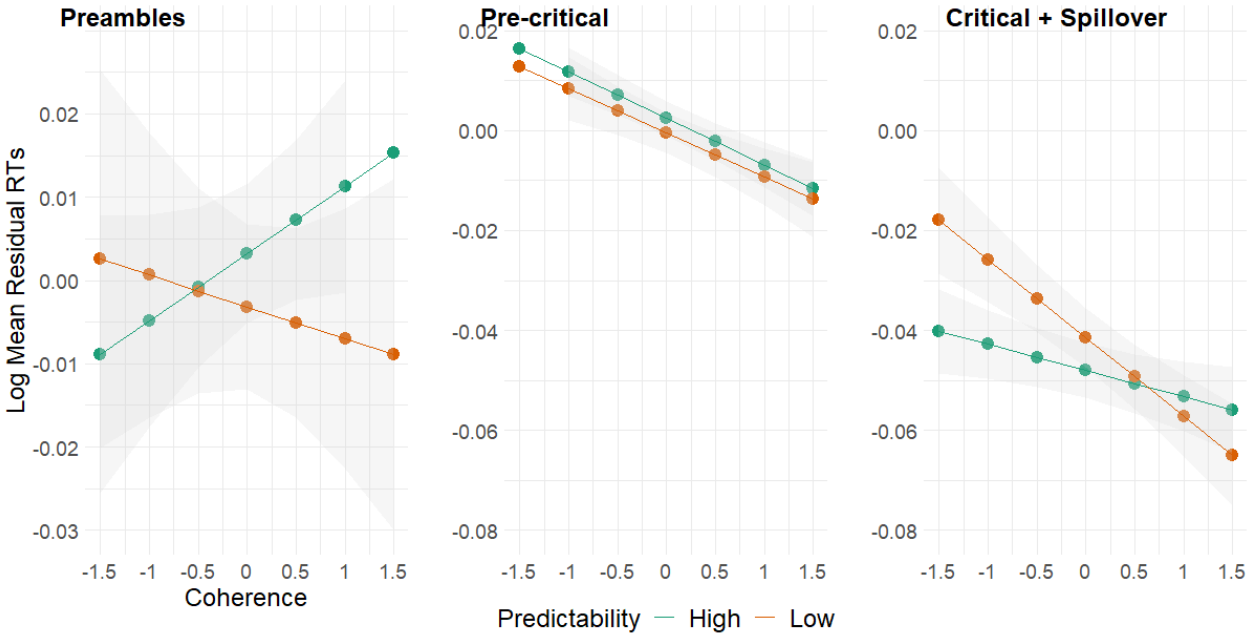

Figure 2.3. Experiment 1 exploratory analysis results; shaded areas depict the standard error.

Critical Word. When considering the critical word only, we found that critical words preceded by more coherent preamble contexts had a processing advantage (β= -0.009, SE= 0.004, *t*= -2.22, *p*= 0.03). However, there was no facilitation of highly predictable critical words (β= -0.004, SE= 0.003, *t*= -1.55, *p*= 0.12). Reading times were not affected by the Zipfian frequency of the critical word (β= -0.003, SE= 0.003, *t*= -0.86, *p*= 0.39), nor by an interaction between coherence and predictability (β= 0.006, SE= 0.005, *t*= 1.24, *p*= 0.22).

2.3.3. Discussion

Experiment 1 found that people were faster to read target sentences when they occurred after a highly coherent context. This effect emerged at the pre-critical ROI, which is where the lack of coherence between the preambles and target sentence would begin to become apparent, and (in our improved analysis pipeline) extended into

the critical word and spillover ROI. Thus, it appears that contextual support from the global discourse impacted reading times to a greater degree than contextual support from the local sentence in this experiment.

The absence of a significant predictability effect at the critical word and spillover ROI is also notable. We identify three possible reasons as to why this effect was weaker than expected. First, with the addition of a global context, it is possible that information from this takes precedence over information from the local (sentence) context, as suggested by some previous studies (Hess et al., 1995; Ledoux et al., 2006). Thus, local predictability may have little influence in this situation. An alternative explanation is that the wider experimental context discourages predictive processing due to the frequent occurrence of unexpected stimuli: 75% of trials in the experiment contained at least one kind of expectation violation (either a lack of coherence between preamble and target sentence, the presence of a less predicted critical word, or both). This conclusion would be consistent with Brothers et al.'s (2017) proposal that participants reduce their use of predictive processing when in an experimental environment in which expectations are frequently disconfirmed. Finally, because we used low predictable target words that were nevertheless semantically plausible for their preceding context, the predictability effect may be small and thus we may have been unable to detect it.

We found no hint of an interaction effect between coherence and predictability. This could indicate that local expectations are not influenced by the presence of discourse-level incoherence, which would suggest a more automatic mechanism for predictive processing. However, it is difficult to conclude this without observing a reliable main effect of predictability. Thus, we designed Experiments 2 and 3 to test whether an interaction effect would be present in conditions that promoted larger effects of predictability. In Experiment 2, we altered the ratio of high and less coherent contexts across the experiment to a 75:25 split, in order to reduce the overall frequency with which expectations were violated and address this potential explanation for the absence of the predictability effect. In Experiment 3, we replaced the low predictability critical words (which, though not predicted, were semantically plausible) with anomalous words that violated the semantic expectations provided by the target sentence. This was intended to induce the strongest possible predictability effect, thereby addressing another possible explanation for its absence.

2.4. Experiment 2

The purpose of Experiment 2 was to address the possibility that the lack of predictability effects in Experiment 1 was due to the even ratio of more and less coherent contexts. It is possible the high prevalence of incoherent narratives led participants to disengage from predictive processing during the experiment as a whole (Brothers et al., 2017). Incoherent passages are relatively infrequent in natural language contexts, with interlocutors typically maximising the amount of coherence in their message (Black, 1988; Grice, 1975, 1989). Thus, to determine whether the same results would be found under conditions where topic shifts occurred less frequently, we reduced the frequency of less coherent passages from 50% to 25% of trials.

2.4.1. Methods

Participants

Participants were recruited through the Testable Minds online participant pool. We required that the participants were based in the UK, had an approval rating greater than 90%, were native speakers of English and had a maximum age of 40. They were reimbursed $5.60 for their time. In total, we collected data from 134 participants (mean age: 28, range: 18-40, SD = 6.09). 26 participants were excluded for failing to meet our pre-registered performance criteria (see below), leaving 108 participants in the final analysis.

Stimuli

The stimuli used in Experiment 2 were identical to Experiment 1; however, we altered the ratio of trials with more and less coherent preambles in order to reduce the amount of topic shifts that occurred across the experiment as a whole. This ratio was changed to 75:25 for more and less coherent contexts. To do so, we took the experimental lists from Experiment 1 and replaced half of the LC trials with their equivalent HC preambles. We ensured that each block (every 40 trials) had the same ratio of HC and LC trials (15 for each HC and 5 for each LC, respectively). We also ensured that each trial appeared in all four conditions across the experimental lists.

Procedure

The procedural steps for Experiment 2 were identical to Experiment 1, except for one change to the presentation of the preambles. We ensured that preambles appeared on-screen for a minimum of two seconds before advancing to the target sentence, irrespective of how quickly participants responded (unlike in Experiment 1 where participants could advance in less than 2s). This was to encourage participants to process the preambles more fully. Because of this, we did not log transform RTs for this ROI (unlike the exploratory analyses in Experiment 1).

Analysis

As with Experiment 1, we pre-registered our analyses prior to data collection (https://osf.io/zu76h/?view_only=20e5c5e18323495a940e237fd8496386). Our pre-registered analyses for Experiment 2 used the analysis approach from our exploratory analyses of Experiment 1 (except for log-transforming RTs in the preambles ROI). In terms of exclusion criteria, participants were excluded if their comprehension accuracy was lower than 80% (n=21), and if the SDs of their preamble RTs was more than 15s (n=5). Trials with incorrect answers to the comprehension questions were also removed (n=440).

2.4.2. Results

Pre-registered Analyses

Preambles. For the preambles (see Figure 2.3), we found no differences in reading times for more and less coherent conditions (β= -29.65, SE= 54.70, *t*= -0.54, *p*= 0.59), no differences in reading times between high and less predictable conditions (β= -36.02, SE= 37.35, *t*= -0.96, *p*= 0.34), nor any interaction (β=-30.40, SE= 51.89, *t*= -0.59, *p*= 0.56).

Pre-critical Word. At the pre-critical word region, we found a marginal effect of coherence, such that there was a suggestion that participants may have read more coherent conditions faster than less coherent conditions (β= -0.005, SE= 0.003, *t*= -1.74, *p*= 0.08). As expected, we found no impact of predictability on reading times (β= -0.001, SE= 0.002, *t*= -0.73, *p*= 0.46), nor an interaction between the two (β=-0.001, SE= 0.003, *t*= -0.42, *p*= 0.68).

Critical Word and Spillover. At our main ROI, we found no influence of coherence (β= -0.01, SE= 0.01, *t*= -1.46, *p*= 0.15), predictability (β= -0.001, SE= 0.004, *t*= 0.25, *p*= 0.80), or their interaction (β= -0.005, SE= 0.009, *t*= -0.56, *p*= 0.58). Further, there was no effect of critical word frequency (β= 0.003, SE= 0.004, *t*= 0.79, *p*= 0.43).

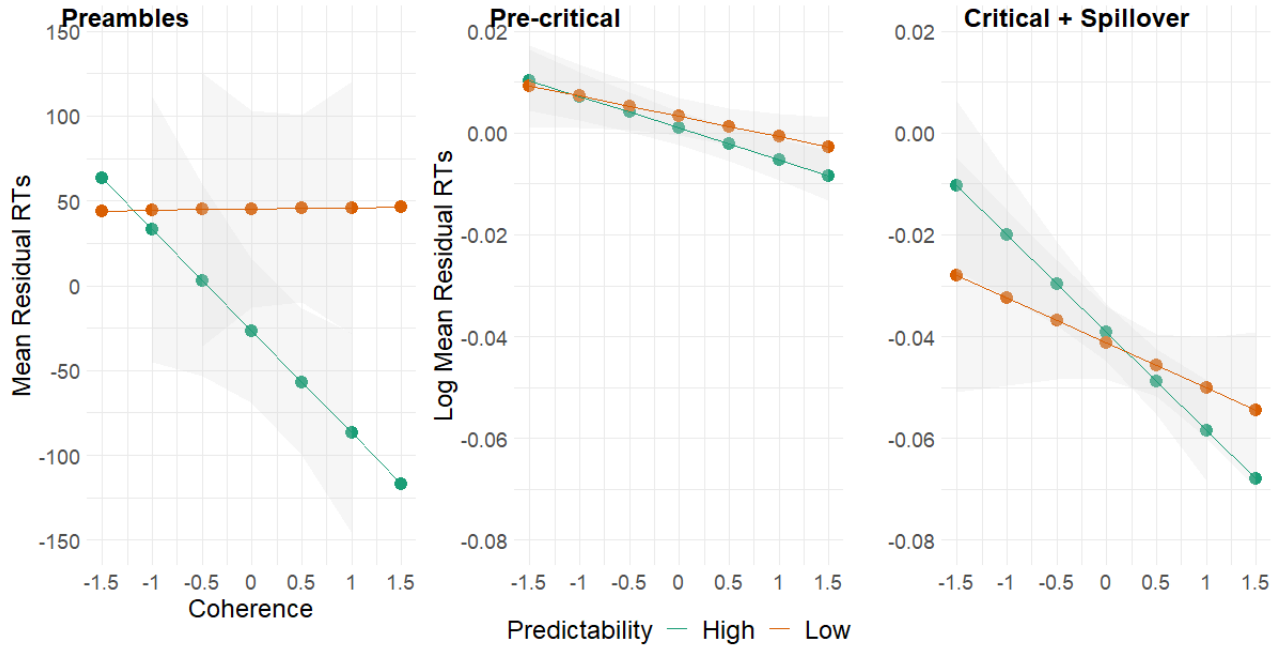

Figure 2.4. Experiment 2 pre-registered results; shaded areas denotes standard error.

Critical Word. At the critical word alone, we found no differences in reading times between more and less coherent conditions (β= -0.01, SE= 0.009, *t*= -1.38, *p*= 0.17), high and less predictable conditions (β= -0.001, SE= 0.003, *t*= -0.42, *p*= 0.67), nor their interaction (β= -0.002, SE= 0.008, *t*= -0.24, *p*= 0.81). Further, there were no differences in reading times for high and less frequent critical words (β= -0.001, SE= 0.004, *t*= -0.32, *p*= 0.75).

Combined Experiment 1 and 2

In order to directly compare the effects from Experiments 1 and 2, we combined the datasets and ran additional linear mixed-effects models, including fixed effects of experiment and its interactions with coherence and predictability.

Preambles. For the preambles region, there was no effect of experiment or an interaction between experiment and the other predictors (all $p > 0.05$).

Pre-critical Word. For this ROI, we found no effect of experiment and there was no interaction with the other predictors (all $p > 0.05$).

Critical Word and Spillover. At our main ROI, we found a marginal effect of coherence (β= -0.009, SE= 0.005, *t*= -1.78, *p*= 0.08). There was no effect of predictability (β= -0.002, SE= 0.002, *t*= -1.16, *p*= 0.25), no interaction (β= 0.0006, SE= 0.003, *t*= 0.21, *p*= 0.84), and no effect of critical word frequency (β= -0.001, SE= 0.002, *t*= -0.46, *p*= 0.65). Further, there was no effect of experiment or an interaction between experiment and the other predictors (all $p > 0.05$).

Critical Word. At this region, we found marginal effects of coherence (β= -0.009, SE= 0.005, *t*= -1.86, *p*= 0.07), predictability (β= -0.003, SE= 0.002, *t*= -1.72, *p*= 0.09) and log frequency (β= -0.004, SE= 0.002, *t*= -1.79, *p*= 0.07). We found no difference in reading times for the interaction of coherence and predictability, no effect of experiment nor any interactions of experiment with the other predictors (all $p > 0.05$).

2.4.3. Discussion

Experiment 2 attempted to test whether the predictability effect (and as an extension, its interaction with coherence) would occur in a more naturalistic experimental environment, where incoherent passages occurred less frequently. However, we found that even after adjustments to the experimental environment to promote predictive processing, still no predictability effect was found. Further, there was no interaction effect with coherence, replicating our findings from Experiment 1. Effects of coherence at the target sentence (emerging at the pre-critical ROI) did not reach statistical significance in this experiment, though they were of a similar magnitude to those observed in Experiment 1 and a combined analysis indicated that the size of the coherence effects did not differ significantly between experiments. We suggest that the lack of a significant coherence effect here may be due to reduced power as less coherent passages were sampled less frequently. For Experiment 3, we returned to an even ratio of high and low coherence trials and investigated the effect of including semantically anomalous (as well as unpredictable) critical words.

2.5. Experiment 3

Experiment 3 was conducted in order to address the absence of a predictability effect in Experiments 1 and 2. It is possible that an effect of predictability did not appear due to

the nature of our critical words. The less predictable conditions in Experiment 1 and 2 were created using a critical word with a lower predictability and plausibility than the target critical word (using the data from Peelle et al., 2020), but was not impossible. They therefore represented a violation of expectations at the lexical level but not the semantic level. It is possible that this manipulation was not strong enough to generate a reliable predictability effect, due to the high degree of semantic feature overlap between the less predictable (but semantically plausible) continuation and the expected continuation (Federmeier & Kutas, 1999). As such, in Experiment 3 we replaced less predictable critical words with semantically anomalous words, which disconfirmed expectations at both the lexical and semantic levels. Thus, we expected this manipulation to produce more reliable effects of predictability, increasing our sensitivity to detect an interaction of this effect with coherence.

2.5.1. Methods

Participants

Participants were recruited through the online platform, Prolific. We filtered the participant pool using the following eligibility criteria: 1) native speakers of English, 2) currently located within the UK, 3) a maximum age of 40, and 4) a minimum approval rating of 90%. Participants were reimbursed with £5.63 for their time. In total, we collected data from 126 participants (mean age: 28, range: 18-40, SD: 5.75). Our final sample size consisted of 97 participants (n=4 removed for a SD of their preambles RTs larger than 15s; n=25 removed for having a comprehension accuracy lower than 80%).

Stimuli

We created two new conditions for each trial that included an anomalous critical word. We replaced the less predictable stimuli with these anomalous stimuli. An example trial with all four conditions can be found in Table 2.3.

Table 2.3. Example sentences across the four conditions in Experiment 3.

Condition	Context Sentences	Target Sentence
High Coherence, High Predictability	Crime was a problem. Residents were starting to feel unsafe.	The police arrested the local gang for selling **drugs** on the corner.
High Coherence, Anomalous		The police arrested the local gang for selling **words** on the corner.
Low Coherence, High Predictability	The city was a safe place. It had a low crime rate and high employment.	The police arrested the local gang for selling **drugs** on the corner.

Low Coherence, Anomalous		The police arrested the local gang for selling **words** on the corner.

We ensured that the anomalous critical words had the same part-of-speech as the high and low predictable critical words, ensuring that the violation would be purely lexico-semantic. Moreover, we matched the anomalous critical words on length (t= 0, p= 1), Zipfian frequency (t= 0.29, p= 0.77) and letter bigram frequency (t= -1.88, p= 0.06) with the highly predictable critical words (see Table 2.4). All of the anomalous critical words had a cloze value of 0 within the local context of the target sentence (t= -165.04, p<0.001). We used the same high and low coherent preambles as Experiments 1 and 2, therefore the matching on preamble length was the same (t= 1.11, p= 0.27).

We did not obtain new coherence ratings for passages containing anomalous words as we assumed that the presence of a strong semantic violation in the target sentence would lead raters to give a low coherence rating to all of these stimuli, irrespective of how coherent the discourse was prior to the violation. This would be problematic as we were interested in how the coherence of the discourse leading up to the critical word influences its processing. Instead, we took the coherence ratings from HC/HP and LC/HP versions of the stimuli and used these as the coherence values in the HC/Anom and LC/Anom conditions. This means that HC/HP and HC/Anom have the same coherence values in our analyses, reflecting the fact that these conditions are identical up to the critical word (and the same for LC/HP and LC/Anom). As in Experiments 1 and 2, coherence differed significantly between HC and LC stimuli (F= 1238.19, p= 2.19×10^{-151}).

Table 2.4. Average psycholinguistic properties of conditions in Experiment 3.

Property	**HC/HP**	**HC/Anom**	**LC/HP**	**LC/Anom**
Coherence	3.83 (0.46)	3.83 (0.46)	2.46 (0.52)	2.46 (0.52)
Preamble Length	13.54 (2.05)	13.54 (2.05)	13.78 (1.92)	13.78 (1.92)
Critical Word Length	4.35 (1.16)	4.35 (0.90)	4.35 (1.16)	4.35 (0.90)
Critical Word Zipf Frequency	4.53 (0.80)	4.56 (0.68)	4.53 (0.80)	4.56 (0.68)
Critical Word Letter Bigram Frequency	1794.57 (1536.06)	1504.95 (1087.85)	1794.57 (1536.06)	1504.95 (1087.85)
Cloze Value	86.23 (6.59)	0.00 (0.00)	86.23 (6.59)	0.00 (0.00)

Note. Standard deviations are shown in parentheses. HC = high coherence; HP = high predictability; LC = low coherence; Anom = anomalous predictability.

Procedure

The procedural steps for Experiment 3 were identical to Experiment 2, except that we returned to presenting each participant with an equal number of HC and LC trials.

Analysis

Similar to Experiments 1 and 2, we pre-registered our analyses prior to data collection (https://osf.io/cqr7a/?view_only=de8f630c89d54c11972017ad994f9c76). All of our analytical steps and exclusion criteria are identical to Experiment 2.

2.5.2. Results

Pre-registered Analyses

Preambles. Model estimates of reading times are presented in Figure 2.4. For the preambles, as expected, we found no differences in reading times for the effect of coherence (β= -36.05, SE= 72.97, *t*= -0.49, *p*= 0.62), no differences in reading times for high and anomalous predictable conditions (β= 46.10, SE= 36.74, *t*= 1.26, *p*= 0.21), nor any interaction (β= -82.74, SE= 57.97, *t*= -1.43, *p*= 0.15).

Pre-critical Word. At the pre-critical word region, we found no effect of coherence (β= -0.00003, SE= 0.003, *t*= -0.01, *p*= 0.99), no differences between high and anomalous predictable conditions (β= -0.001, SE= 0.002, *t*= -0.82, *p*= 0.41), nor any evidence of an interaction (β= 0.004, SE= 0.003, *t*= 1.33, *p*= 0.18).

Critical Word and Spillover. At our main ROI, we found a significant effect of predictability, such that participants read the critical word and spillover region faster for highly predictable critical words than for anomalous critical words (β= 0.005, SE= 0.002, *t*= 2.25, *p*= 0.02). We also found a significant interaction between coherence and predictability (β= 0.009, SE= 0.004, *t*= 2.13, *p*= 0.04), such that the facilitation for highly predictable words was greater on more coherent trials, than for anomalous words. In post-hoc analyses, we ran separate linear mixed-effects models for the high coherence and low coherence trials. A predictability effect was present for the more coherent trials (β= 0.007, SE= 0.003, *t*= 2.18, *p*= 0.03), but not for the less coherent trials (β= 0.004, SE= 0.003, *t*= 1.26, *p*= 0.20). Finally, we found no main effect of coherence (β= 0.003, SE= 0.004, *t*= 0.60, *p*= 0.55), nor an effect of frequency (β= -0.004, SE= 0.003, *t*= -1.39, *p*= 0.16).

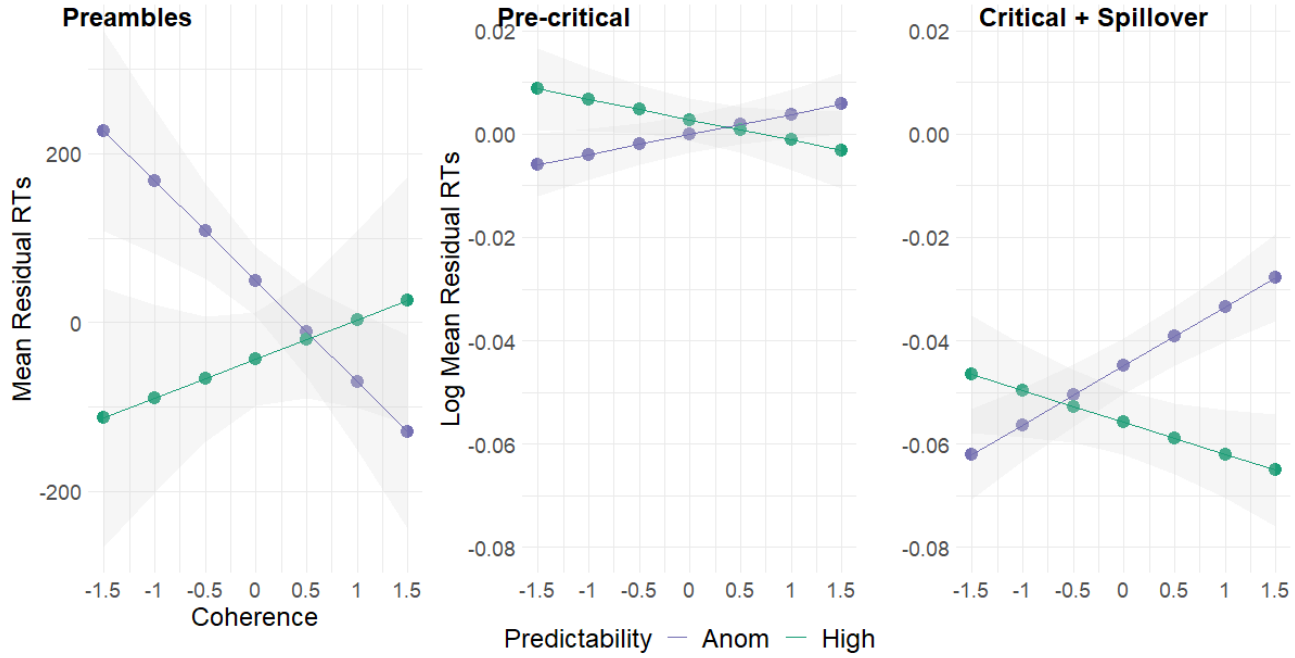

Figure 2.5. Experiment 3 pre-registered results; shaded areas indicate the standard error.

Critical Word. Analyses of the critical word alone produced similar results to those that included the spillover region. There were significant reading time differences between the highly predictable and anomalous conditions (β= 0.005, SE= 0.002, *t*= 2.10, *p*= 0.04). Further, the interaction of coherence and predictability was marginally significant (β= 0.008, SE= 0.004, *t*= 1.91, *p*= 0.06), as was the impact of critical word frequency (β= -0.005, SE= 0.003, *t*= -1.77, *p*= 0.08). No differences in reading times were found for the high and low coherent conditions (β= 0.003, SE= 0.004, *t*= 0.59, *p*= 0.56).

Combined Experiment 1 and 3

To compare the results from Experiments 1 and 3, we also ran further analyses on a combined dataset from these experiments. Analyses included fixed effects of experiment and its interactions with coherence and predictability.

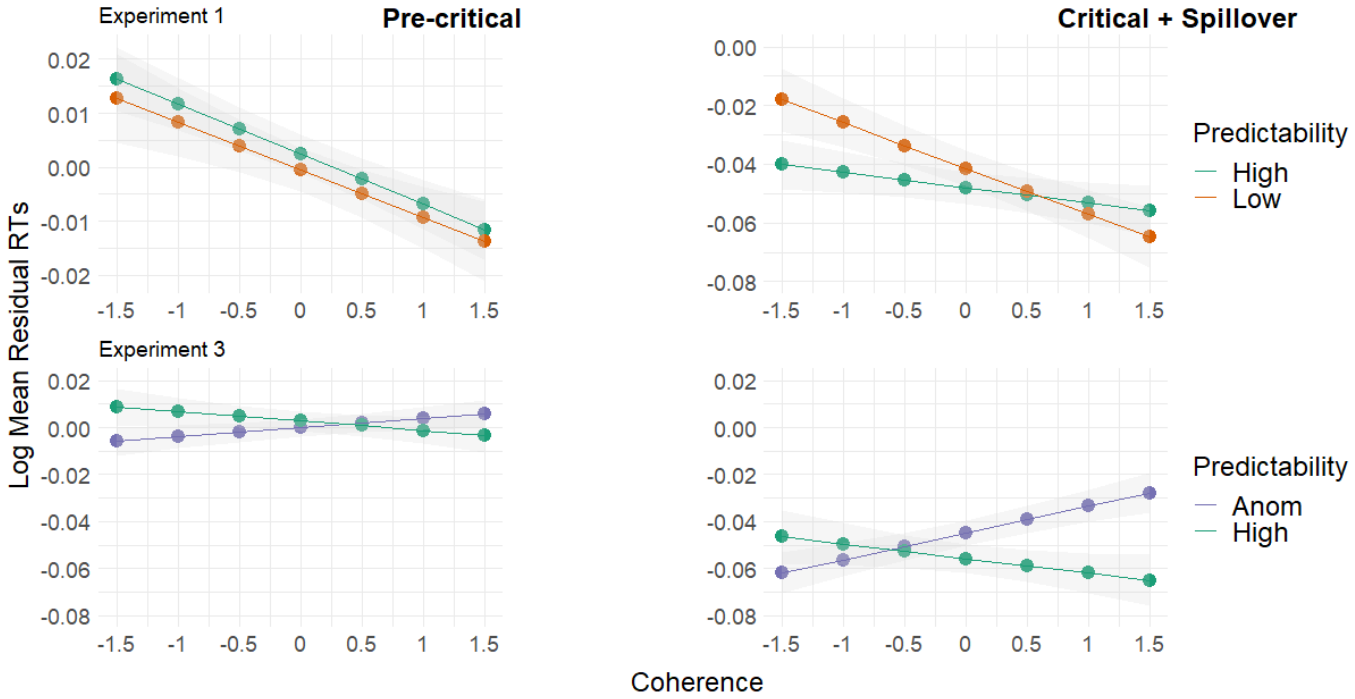

Figure 2.6. Results for pre-critical word and critical word and spillover regions for Experiment 1 (top) and Experiment 3 (bottom); shaded areas denote standard error.

Preambles. At this ROI, we found a significant interaction of predictability and experiment (β= 0.006, SE= 0.003, *t*= 2.06, *p*= 0.04) (see Figure 7.1 in Appendix). This is likely to be a false positive as the low and high predictability trials had the same preambles across both experiments. No other results reached significance, including the test of experiment and its interactions (all $p > 0.05$).

Pre-critical. For the pre-critical region, less coherent trials were read more slowly (β= -0.004, SE= 0.002, *t*= -2.07, *p*= 0.04), and this effect interacted with experiment (β= -0.004, SE= 0.002, *t*= -2.15, *p*= 0.03), such that the coherence effect was larger in Experiment 1 (as shown in Figure 2.5). This is likely due to the fact that a coherence effect was observable in Experiment 1, and not in Experiment 3. There were no other main effects or interactions (all $p > 0.05$).

Critical Word and Spillover. For our main ROI, we found that anomalous/less predictable words were read slower (β= -0.005, SE= 0.002, *t*= -2.76, *p*= 0.006) than highly predictable words and critical words with higher frequency were read faster (β= -0.004, SE= 0.002, *t*= -2.02, *p*= 0.04). Moreover, there was an interaction between coherence and experiment (β= -0.007, SE= 0.003, *t*= -2.25, *p*= 0.02). Coherence had a larger overall impact on reading times in Experiment 1 than in Experiment 3, as an overall coherence effect was not observed in Experiment 3. Importantly, a three-way interaction emerged between coherence, predictability and experiment (β= 0.007, SE= 0.003, *t*= 2.53, *p*= 0.01). This indicates that the pattern observed in Experiment 3, whereby the predictability of the critical word influenced processing only in more coherent passages, differed significantly from that of Experiment 1 (see Figure 2.5 for illustration). There were no other main effects or interactions (all $p > 0.05$).

Critical Word. At the critical word alone, we found that anomalous/less predictable words were read more slowly than highly predictable words (β= -0.005, SE= 0.002, *t*= -2.93, *p*= 0.004), as well as that critical words with higher log frequency were read faster than those with lower log frequency (β= -0.007, SE= 0.002, *t*= -3.16, *p*= 0.001). Further, a three-way interaction between coherence, predictability and experiment emerged (β= 0.006, SE= 0.003, *t*= 2.13, *p*= 0.03), and a marginal interaction between coherence and experiment was present (β= -0.005, SE= 0.003, *t*= -1.81, *p*= 0.07). We found no other main effects or interactions (all $p > 0.05$).

2.5.3. Discussion

In Experiment 3, we replaced the less predictable critical words with semantically anomalous words that violated readers' expectations at the semantic as well as the lexical level. We expected to observe larger effects associated with the critical word, and we did: there was a main effect of predictability, whereby people were slower to read anomalous critical words, compared with highly predictable critical words. However, the size of this effect varied according to discourse coherence, with anomalous critical words slowing processing only when preceded by more coherent contexts. In other words, the more coherent global context facilitated the processing of highly predictable critical words but impaired the processing of anomalous critical words. We interpret these results as evidence of variation in the engagement of predictive processes. Reading of highly predictable critical words was facilitated when these were preceded by a highly coherent context, suggesting that these conditions encouraged reliance on expectations. At the same time, reading of anomalous words was slower in highly coherent contexts, consistent with the need to overcome the effect of strong but incorrect expectations. Thus, these results suggest that the degree to which predictive processing is engaged depends on the surrounding context. It is possible that comprehenders downregulate predictive processing when the global context is not informative, thus reducing the effects at the level of the local sentence.

Another interpretation of the interaction effect is that it relates to variations in how long readers spend verifying that a word is anomalous. Specifically, it is possible that more time is taken to confirm whether a word is semantically anomalous when it is embedded in a more coherent discourse, as the passage made sense until that point. In other words, the anomaly may be more surprising on high coherence trials. It is difficult to disentangle this possibility from a predictive processing account, as both are rooted in readers' expectations about the critical word. In addition, an explanation based solely on anomaly processing does not provide an explanation for facilitation of predictable words, which we also observed.

2.6. General Discussion

The current study addressed how discourse coherence influences predictive processing during language comprehension. We were particularly interested in how the validity of the global discourse context affected comprehenders' use of local sentence constraints. To this end, we conducted three online, self-paced reading experiments

using three-sentence discourses. Our stimuli were manipulated on both the predictability of the critical word given its local sentence, and the coherence of the target sentence, given the preceding preamble. Results from Experiment 1 suggested that global discourse had an influence on processing, with a coherence effect emerging at the pre-critical ROI and extending into the critical word and spillover ROI. This suggests that information from the global discourse context is used to facilitate comprehension. However, in Experiment 1, while discourse coherence benefited processing, no local predictability effects were observed. In our second experiment, we used a lower proportion of less coherent trials but found similar results to Experiment 1. Experiment 3 used anomalous critical words that more strongly violated comprehenders' expectations. Here, an effect of predictability did emerge, with facilitation for highly predictable words compared to anomalous words. Importantly, however, this effect was only observed when the target sentence was embedded in a highly coherent passage, with no predictability advantage present for less coherent discourse. These results indicate, first, that relatively subtle manipulations of discourse coherence influence the speed of comprehension and second, that a reduction in the coherence of a discourse appears to reduce the effects of local sentence constraints. This latter effect is in line with the proposal that comprehenders use a range of environmental cues to regulate their use of predictive processing. We will discuss each of these effects in turn.

2.6.1. Influence of Coherence

In our first two experiments, the impact of coherence on how comprehenders process upcoming material was clear. Trials preceded by more coherent preamble contexts were read faster than those preceded by less coherent contexts. In Experiment 1, this coherence effect emerged at the pre-critical region and extended into the critical word and spillover region. The emergence of the coherence effect at the pre-critical region indicates the rapid influence of cues from global context on reading times. The pre-critical ROI is the first region where the lack of coherence between the preamble context and the target sentence could be detected. While the coherence effect in Experiment 2 did not reach statistical significance, possibly due to the lower sampling of less coherent contexts, the numerical effect was of a similar magnitude as in Experiment 1. These results hint at the privileged position information from the global context plays in reading comprehension. Our results are consistent with previous findings that the presence of a highly coherent discourse context facilitates comprehension, while deviation from the existing discourse model results in a slowdown (Albrecht & O'Brien, 1993; Hess et al., 1995; Myers & O'Brien, 1998; O'Brien & Albrecht, 1992; Schwanenflugel & White, 1991; Stewart et al., 2009).

We theorise that comprehenders initially build a discourse model of the events portrayed in the preamble contexts, then when faced with a target sentence that is not coherent with this established mental model, processing difficulties emerge. The mismatch between the established mental model and the incoming incoherent information can be established early into the target sentence and, indeed, we found that the coherence effect emerges at the pre-critical region. These processing difficulties appear to be sustained in time, such that they impact the processing of later

words within the target sentence. This suggests that the comprehender experiences continued difficulties in integrating the incoming information into their established mental model of the discourse. The prolonged slowing in reading times could be indicative of extended and perhaps repeated attempts to re-update the mental model as new information becomes available (Albrecht & O'Brien, 1993; Van Petten & Luka, 2012). One interesting line of future research would be to have a finer-grained measure of when, and for how long, the coherence effect impacts processing. One way in which this could be achieved would be to track the emergence of coherence effects at the single word level, using either RTs or ERP measures, to determine when disruptions to coherence become apparent to comprehenders. Predictions from NLP models (Frank et al., 2013; Szewczyk & Federmeier, 2022) could inform this effort by providing information on how breaks in coherence influence the surprisal associated with words in the target sentence. An alternative line of inquiry could also track participant's behaviour over the experiment to explore whether they become aware of the high uncertainty of coherence and adjust their processing strategies accordingly.

Surprisingly, in our third experiment, we did not find an effect of coherence. This was unexpected as the stimuli were the same as those presented in Experiment 1 up until the anomalous critical word. As such, we would have expected a similar coherence effect to emerge at the pre-critical ROI across all experiments. The fact that this is absent in Experiment 3 suggests that a difference in the experimental environments (between Experiments 1 and 3) may have induced changes in processing strategy. In Experiment 1, unpredictable critical words were less plausible than the high predictable critical words but not impossible, so they could be accommodated by updating the discourse model. In contrast, Experiment 3 used semantically anomalous critical words that were entirely incompatible with any meaningful interpretation of the discourse. It is possible that participants became aware that on many trials they would encounter an anomalous critical word, and their discourse model would fail. This could have led participants to delay integration of the target sentence with the preambles until later in the sentence to avoid wasted effort. Our current findings are not able to concretely conclude this, so further study on this is warranted.

2.6.2. The Role of Predictability

The first two experiments failed to observe a main effect of predictability. In these experiments, we created our less predictable critical word condition by swapping the highly predictable critical word with a lower cloze continuation of a matching syntactic role so that it still remained plausible. Thus, while the specific lexical form of our less predictable critical words was not expected, they were still compatible with expectations at the syntactic and semantic levels and may have been facilitated to some extent (Amsel et al., 2015; Federmeier & Kutas, 1999; Metusalem et al., 2012; Wlotko & Federmeier, 2012). To address this, we used semantically anomalous critical words in Experiment 3 (replacing the less predictable condition). Here, we found that when coherence was high, comprehenders were faster to read the highly predictable continuations than the anomalous targets. This was as expected: all of the target sentences allowed comprehenders to generate a strong expectation about the critical word, and processing of this word was then facilitated when it appeared. However, if a

semantically anomalous word was observed instead, processing became more difficult as the prior expectation was unhelpful (and disconfirmed). Critically, however, this predictability effect did not occur when the target sentence appeared in a less coherent discourse. On these trials, the contents of the target sentence were the same, and therefore in principle, would still lead to an expectation that favours the highly predictable continuation. However, no such facilitation was observed. In addition, reading of anomalous words was faster when they appeared in less coherent passages, suggesting that no prediction was disconfirmed when coherence was low. Thus, it appears that when the preamble contexts were less coherent, local sentential constraints no longer affected processing to the same degree as for more coherent contexts.

These results add to the evidence that expectations about upcoming language material are recruited flexibly and are influenced by a range of contextual cues beyond the immediate sentence (Brothers et al., 2020; Huettig, 2015). For example, previous studies have shown that the degree of facilitation for predictable words depends on the comprehender's knowledge of the speaker and the frequency with which predictions are violated in the experiment (Brothers et al., 2017, 2019). Thus, when people have reason to believe that their predictions will not be valid, they appear to pre-activate predictable words less. The present results indicate that the coherence of the current discourse is another cue that comprehenders use to regulate their use of predictive processing. On less coherent trials, we propose that as participants processed the target sentence, they became aware that the mental model established during the preamble was no longer valid. The introduction of new information that was incompatible with the existing mental model may have acted as a signal that expectations were likely to be unreliable. Thus, the language system disengaged attempts to use expectations to facilitate processing. This would explain the absence of an advantage for highly predictable continuations over semantically anomalous words.

Importantly, our effects occurred even though the target sentence itself was still highly constraining towards a particular completion. The uncertainty regarding the state of the global discourse seemed to prevent these constraints from being used to aid processing. We also highlight that coherence itself was not a reliable cue that the target sentence would contain a violation, since violations were equally common in high and low coherence trials. As such, we do not think these effects reflect a specific strategy used in our experiment. Instead, the downregulation of predictive processing appears to be a natural response to detecting that the discourse is in a state of uncertainty.

Our results are in opposition to those found by Camblin et al. (2007), whereby facilitation from local within-sentence associations were largest when preceded by extremely incoherent contexts. In that study, the incoherent contexts consisted of sentences scrambled between different trials, so they were completely uninformative and no discourse model could be constructed. We suggest that the difference between the current study and Camblin et al. (2007) is that when the global context is completely uninformative, comprehenders disregard it and rely entirely on information at the local level. In the current study, however, the more subtle manipulation of coherence may have led comprehenders to attempt to form a coherent representation of the passages.

In this situation, we suggest that disruption to coherence acts as a signal that local constraints are less likely to be reliable.

2.6.3. Limitations and Future Work

The current study has some limitations. One potential limitation is the use of relatively subtle disruptions to global discourse coherence. This means that our findings are unable to definitively answer how information from the local context is recruited when the global context is completely uninformative. However, our stimuli were designed to contain coherence breaks that are representative of how coherence can be disrupted in natural speech, rather than using artificially meaningless contexts. It could be interesting to investigate how our results compare with less natural disruptions to coherence in future. But it would be important that these types of trials only occur rarely in the study, as there is already evidence that comprehenders are sensitive to the statistical structure of the experiment (Brothers et al., 2017). It is possible that frequent strong coherence violations may lead to a total disengagement in predictive processing.

Another possible suggestion for future work could look at the mechanism behind the processing differences observed for highly predictable critical words, compared to semantically anomalous continuations. One option could be that facilitation occurs for highly predictable words because a correct expectation has been confirmed. However, it is also possible that this processing difference is due to a slowdown in the anomalous condition as expectations have been violated (Van Petten & Luka, 2012). In order to adjudicate between these two possibilities, a neutral condition that contains a low cloze, low constraint target sentence would need to be included.

In conclusion, the current study demonstrates that comprehenders are sensitive to relatively subtle changes in the coherence of a discourse and that processing is slowed when these occur. The advantage of a coherent global context emerged at the earliest possible position and extended in time during subsequent processing. Further, our findings demonstrate that a lack of coherence eliminates within-sentence facilitation for predictable words. This suggests that the presence of a narrative shift serves as a cue to temporarily reduce the reliance on predictive processing. Future work could further investigate the nature of this coherence check through the use of finer-grained temporal measures, such as ERPs.

Chapter 3

Predicting discourse context effects from surprisal

3.1. Abstract

We know that context can facilitate language comprehension. Previous research has shown that discourse coherence influences this contextual facilitation, with comprehenders making stronger predictions about upcoming words when reading highly coherent narratives. However, it is unclear whether the predictions made by Large Language Models (LLMs) exhibit similar discourse-level influences. As such, we investigate whether surprisal values from LLMs reflect longer context effects. We calculated word-level surprisal values (as a measure of prediction strength) for passages that vary in coherence. We used these to predict human reading times for the same passages collected from 289 participants. We found that surprisal only predicted reading times early in the target sentence, and that GPT-2's surprisal values were not influenced by discourse coherence, in contrast to human reading data. This has implications on the use of Transformer-based LLMs in modelling human cognition.

3.2. Introduction

Language comprehension is an incremental process where a meaningful representation is built up over time, involving the integration of multiple sources of linguistic information such as syntax, semantics, pragmatics and discourse (Sedivy et al., 1999). Empirical work across cognitive neuroscience and psycholinguistics has demonstrated a processing advantage for highly predictable words, such that words which are predictable given their context are processed faster than those which are not (Altmann & Mirković, 2009; Federmeier, 2007; Kutas & Hillyard, 1980, 1984). This facilitation effect has been observed for a range of linguistic features, such as gender agreement (Van Berkum et al., 2005; Wicha et al., 2003, 2004), definiteness (Carter & Nieuwland, 2022; Fleur et al., 2020) and world knowledge (Hagoort et al., 2004; Hald et al., 2007; Metusalem et al., 2012). However, the question as to which cognitive mechanisms underpin this processing advantage remains.

Within cognitive modelling and computational linguistics, Surprisal Theory is a popular account of predictability. Surprisal Theory asserts that the processing difficulty of a word is inversely proportional to its expectancy, such that the less expected a word is given its context, the higher its surprisal and therefore the more difficult it is to process (Hale, 2001; Levy, 2008; Venhuizen et al., 2019). Originating in information theory, where it is also referred to as Shannon information, surprisal is quantified as the negative log probability of a word given its preceding context (Shannon, 1948). Another metric which is commonly used to explicitly model the influence of predictability is

entropy, which can be understood as a quantification of the degree of uncertainty surrounding a next-word prediction (Frank, 2013; Lowder et al., 2018). Both metrics have been successfully used to quantify word-by-word predictability during incremental sentence processing (Hale, 2016; Willems et al., 2016). Most commonly, these metrics have been used to model experimental data, such as measures of eye-tracking, reading times (RTs) and N400 amplitudes (Frank et al., 2013; Goodkind & Bicknell, 2018; Michaelov et al., 2023b; Michaelov, Bardolph, et al., 2024; Monsalve et al., 2012). Typically, surprisal values are extracted from artificial neural networks that have been exposed to language and trained to perform a next-word prediction task, similar to a Cloze task, as exemplified by connectionist cognitive models (Elman, 1990; Frank et al., 2019; McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1987; Taylor, 1953). These language models (LMs), as well as those from recent developments in deep learning and NLP, can be understood as inherent models of linguistic prediction that are only based on language input (Michaelov & Bergen, 2020). As such, they serve as a useful tool to explore theoretical assumptions surrounding predictive processing in language.

One branch of research has focused on identifying the precise relationship between word predictability and the observed processing advantage. Szewcyzk and Federmeier (2022) investigated what type of function links word probability and contextual facilitation, as measured through the N400. They detail two views regarding the possible relationship. The first, put forward by Brothers and Kuperberg (2021), is the "proportional preactivation account", which states that words are preactivated in proportion to the statistics of the linguistic environment. Here, contextual facilitation is explained as a match between the preactivated representations and the incoming linguistic stimulus, represented as a linear relationship (DeLong et al., 2005; Schwanenflugel & LaCount, 1988; Staub, 2015; Van Petten & Luka, 2012). The second view assumes that all words in the lexicon are subject to contextual facilitation, maintaining and updating a probability distribution where the amount of update is equivalent to the word's surprisal. This view assumes a logarithmic relationship between human processing data and predictability measures (Aurnhammer & Frank, 2018; N. J. Smith & Levy, 2013). In their study, Szewcyzk and Federmeier (2022) found that the relationship between word predictability and N400 amplitudes is linear for expected words, but logarithmic for unexpected words. Shain and colleagues (2024) similarly evaluated the different theoretical assumptions that can explain the facilitation effect observed. They compared a facilitation view (i.e., linear effect) with a cost view (i.e., logarithmic effect) by modelling RTs with LM surprisal and constraining the mapping between predictors. The authors found consistent evidence for a logarithmic relationship between surprisal and human RTs.

Another branch of research has focused on evaluating which LMs demonstrate the best fit to human data, with the idea that certain architectural designs may better account for aspects of human processing. For example, Goodkind and Bicknell (2018) compared surprisal values from a range of LMs, such as n-gram models and recurrent neural networks (RNNs) with Long-Short Term Memory (LSTMs), finding that the psychological predictive power of surprisal linearly increased with LM linguistic quality, measured using model perplexity. Perplexity is a common evaluation metric for LMs; it

is defined as the exponentiated average negative log-likelihood of a sequence. Intuitively, this relates to how much of the probability mass from the model is assigned to the correct prediction, where a lower perplexity is indicative of a more accurate LM. Wilcox and colleagues (2020) performed a similar evaluation of modern neural networks, comparing computational architectures and training dataset sizes. They found a linear relationship between LM surprisal and RTs across differences in model architecture and training datasets, which is in line with previous findings (Goodkind & Bicknell, 2018; N. J. Smith & Levy, 2013). They also found a generally positive relationship between a model's next-word prediction accuracy, and its ability to predict human RTs, demonstrating that Transformer models (here, GPT-2; Radford et al., 2019) exhibited the best fit to the human data. In addition, Merkx and Frank (2021) compared Transformer-based models and RNNs on their ability to predict human RTs and N400 amplitudes. Transformer-based models benefit from self-attention layers, which can "attend" to the previous input directly (Vaswani et al., 2017). As such, they are generally thought to be cognitively implausible models, whereas RNNs are considered to be more cognitively plausible models (Elman, 1990; Frank et al., 2019; Jordan, 1997). Results showed that Transformer-based models provide a better fit to the data than Gated Recurrent Units (a type of RNN) which the authors state raises questions about how good RNN models are as cognitive models if they are outperformed by a cognitively implausible model.

Much of the research on predictability-based facilitation has focused on single sentences, while work that has modelled this effect using surprisal has often opted for naturalistic datasets, designed to provide broad language coverage (Futrell et al., 2017; A. Kennedy et al., 2003, 2013). This is potentially problematic because multi-sentence contexts make significant contributions to the representation of linguistic content (Brothers et al., 2017; Hagoort et al., 2004; Nieuwland & Van Berkum, 2006; Zwaan, 2016). One recent notable exception to this is work by Michaelov and colleagues (2023a) which provides a modelling account of the results from Nieuwland and van Berkum (2006). Briefly, Nieuwland and van Berkum (2006) found smaller N400 effects for noncanonical phrases such as "the peanut was in love" compared to canonical phrases like "the peanut was salted" when they were preceded by a discourse context that encourages an animate representation of the peanut. The authors suggest this is due to a shift in the situation model (van Dijk, 1999; Zwaan & Radvansky, 1998). Michaelov and colleagues model these N400 effects using LMs that have no explicit situation model and find that, when presented with the same contexts, surprisal for critical words in noncanonical conditions (i.e., "the peanut was in love") was smaller than the surprisal for canonical conditions (i.e., "the peanut was salted"). This study provides an insight into how LM surprisal may trend with discourse-level phenomena. This is because the Nieuwland and Van Berkum (2006) results are typically interpreted as evidence for the influence of discourse-level world models on prediction. These world models are thought to be outside of the language system itself. However, the results from Michaelov and colleagues (2023a) suggest that these effects can be explained by variations in surprisal which operate purely at the lexical level.

In the present study, I investigate how well surprisal values derived from a Large Language Model (LLM) account for RTs when people read multi-sentence passages that

vary in their levels of internal coherence. Previously, I investigated how discourse coherence may impact predictive processing with a focus on subtler shifts in context, rather than outright violations (see Chapter 2). From this, I found that discourse coherence modulates the degree of contextual facilitation, and that comprehenders will make use of cues from both global (i.e., discourse) and local (i.e., sentence) contexts, especially when semantic violations are present at the local sentence. To explore how surprisal maps onto markers of human sentence processing that require cue-based retrieval from different levels of context and their subsequent integration, here I investigate whether LM surprisal reflects longer context effects. I do this by comparing if model predictions are significantly influenced by the same linguistic features that elicited a significant difference in human RTs, and by asking if LM surprisal explains additional variance in predicting these RTs. I aim to shed light on what sorts of contextual facilitation effects surprisal theory can account for.

To do this, I took the self-paced reading (SPR) dataset of coherence and predictability manipulations from Chapter 2, and input the stimuli to GPT-2, extracting next-word probabilities and calculating the surprisal values. I first fit linear mixed effects models (LMEMs) predicting surprisal from the experimental conditions. This provides information on whether surprisal is sensitive to subtle changes in the coherence of discourse passages. I then predicted RTs in the dataset using surprisal and other low-level cognitive predictors known to affect RTs. This allowed me to assess the amount of additional variance in the RT data that can be explained by surprisal, on top of the experimental conditions. I again evaluated the model by comparing goodness-of-fit with a baseline.

3.3. Methods

3.3.1. Data

Stimuli

My stimuli were taken from Chapter 2, which includes three self-paced reading experiments exploring how information at different levels of context impact predictive processing. In this study, the stimuli were created to contain a manipulation of coherence, as a global contextual cue, and a manipulation of predictability, as a local sentence cue. The stimuli (n=160 per condition) consist of three-sentence trials, where the first two sentences are either highly coherent with the following target sentence, or less coherent. The high coherent preambles were made to form a meaningful context, whereas the low coherent preambles were designed such that the target sentence would instigate a topic shift (Zwaan & Radvansky, 1998). The coherence manipulation was verified in a prior norming study. The same syntactic structure was used for both high and low coherence preambles to minimise changes in processing demands downstream. The target sentences were adapted from Peelle et al. (2020), who provided completion norms for English sentences. I extracted sentences with high cloze completions (> 0.75) to become my high predictable conditions. I then created my low predictable conditions by substituting the highly predictable critical word with a

different lexical item that was plausible. I also created an anomalous condition for the final experiment with a semantically anomalous critical word in order to strengthen the predictability manipulation. I defined three regions-of-interest for my trials, which can be found in Figure 3.1. As these stimuli were used in a SPR study, I added an additional spillover region to account for the delayed nature of SPR as a measure of online language comprehension (Jegerski, 2013).

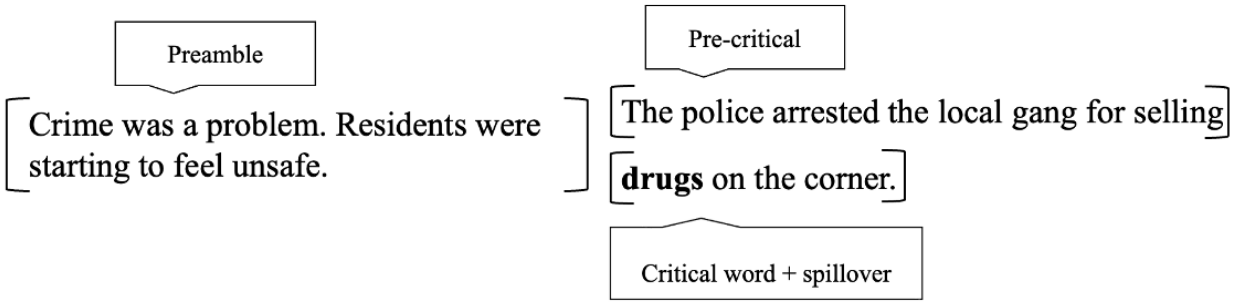

Figure 3.1. Example trial with regions-of-interest from SPR experiments.

Self-paced reading data

I collected RTs across three online SPR experiments using the stimuli described above. Participants saw the two-sentence preambles onscreen in full, followed by the target sentence presented word-by-word. They controlled the length of time the stimuli were presented on screen through keypress. Three different experiments were conducted. Experiment 1 had a 2x2 factorial design of my coherence and predictability manipulations to answer whether coherence impacts downstream predictive processing. Experiment 2 used the same design, but featured a 75:25 ratio split of high and low coherent trials to encourage participants to rely on predictive processing, given prior evidence that the statistical environment of the experiment can impact how likely comprehenders engage in predictive processing (Brothers et al., 2017). In Experiment 3, I replaced the low predictability condition with the anomalous condition to further investigate the limits of predictive processing by strengthening the predictability manipulation. The previous results are as follows:

- Experiment 1: participants were faster to read target sentences preceded by a highly coherent context. This effect emerged at the pre-critical ROI and extended throughout the sentence.
- Experiment 2: a similar coherence effect was observed.
- Experiment 3: participants were faster to read highly predictable critical words than anomalous critical words. Further, anomalous critical words slowed reading to a greater extent when preceded by highly coherent contexts.

3.3.2. Analysis

The language models were run using the *transformers* (Wolf et al., 2020) library in Python using *Pytorch* (Paszke et al., 2019). All analyses were run in R using *RStudio* (Posit Team, 2024), and made use of the *lme4* (Bates et al., 2014), *lmerTest* (Kuznetsova et al., 2017), and *tidyverse* (Wickham et al., 2019) packages.

Extracting surprisal estimates

To calculate the surprisal values, I input my experimental stimuli to GPT-2, a pre-trained, autoregressive LLM (Radford et al., 2019). The training objective for the model is next-word prediction, whereby a model has to predict the next word based on the previous context, which is similar in nature to the Cloze task (Taylor, 1953). I accessed the model through the Huggingface API[1]. GPT-2 was trained on WebText[2], which was tokenized using Byte Pair Encoding (BPE) and has a vocabulary size of 50,257. To run the experimental stimuli through the model, I presented the stimuli for each trial incrementally, each time inputting N words and extracting the probabilities of word N+1. As the pre-trained tokenizer is based on the BPE algorithm, some words were tokenized at the subword level. In these cases, I extracted the next-bpe probability and appended the subword level token to the context to ensure that the same incremental approach was applied at the subword level. To calculate the surprisal from the extracted probabilities, I took the negative log of these probabilities.

Surprisal-based linear mixed-effects models

In my investigation of whether coherence and predictability influence surprisal, I focused my analyses on effects within the target sentence, as this is the location where the experimental manipulations would be predicted to impact processing. I split the target sentences into two regions-of-interest around the critical word. These ROIs are shown in Figure 3.2. To emphasise, although my LMEMs only focus on the target sentence, I ran the full trials through GPT-2 and so the surprisal values reflect the negative log probability of each word, given all previous words in the preambles and target sentence, combined. As such, even though the linguistic material presented in the pre-critical ROI is identical across all conditions, the surprisal values will differ based on the coherence condition (i.e., the preamble context it was paired with). Surprisal values were averaged over all of the words contained in each ROI.

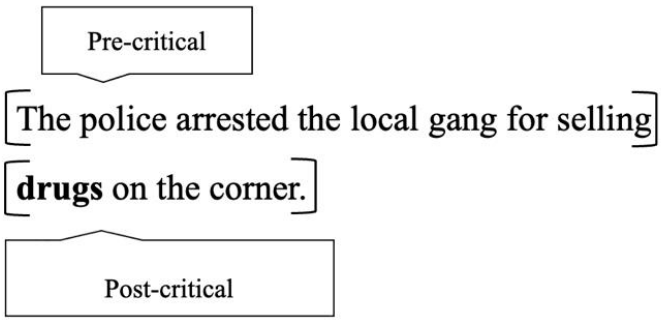

Figure 3.2. Example trial with regions-of-interest for surprisal and RT LMEMs.

I fit linear mixed effects models predicting the average surprisal of each ROI (items: n = 960). I ran a baseline model that predicted average surprisal from a random

[1] https://huggingface.co/openai-community/gpt2
[2] https://paperswithcode.com/dataset/webtext

slope for items alone, and a full model that contained coherence and predictability as main effects, alongside their interaction, with the same random effects structure.

RT-based linear mixed-effects models

To predict the RTs from Chapter 2, I adapted some of the pre-processing steps reported in that chapter. For each experiment, the ROIs underwent winsorisation of the RTs that occurred outside the range of 2 SDs from the mean, calculated for each participant. I fit LMEMs on the log RTs from the three experiments, separately (items: n = 960; exp 1 participants: n = 99; exp 2: n = 98; exp 3: n = 97). All categorical variables were factorised using contrast coding. Moreover, here I included the log RT for each word as a separate observation, rather than averaging over the ROI as in Chapter 2. The baseline models predicted log RT with coherence and predictability as main effects, alongside their interaction. Low-level linguistic features known to impact reading were also included as covariates. These were the scaled word length and the log-transformed Zipfian frequency, derived from SUBTLEX-UK (van Heuven et al., 2014). I additionally included by-subject and by-item random slopes. The full models included surprisal as an additional main effect, keeping all other predictors the same.

3.3.3. Evaluation

To evaluate the goodness-of-fit between the baseline models and full models, I used a Type III ANOVA with Satterthwaite's method for estimating degrees of freedom. I report the χ^2, degrees of freedom and p-values, where significance is defined as $p < 0.05$. I also report the summary statistics and associated p-values of the LMEMs with surprisal.

3.4. Results

3.4.1. Predicting surprisal from experimental conditions

Estimated effects of the model predicting surprisal values from experimental conditions are shown in Figure 3.3. When predicting surprisal from coherence and predictability, I found a significant effect of coherence in the pre-critical ROI, such that as the coherence increases, the average surprisal decreases ($\beta = -0.09$, $SE = 0.02$, $t = -5.19$, $p < 0.001$). On trials where raters rated the stimuli as being highly coherent, surprisal values were lower, indicating that GPT-2 was better able to predict the content of the target sentence. There was no significant effect of predictability (hp vs anom: $\beta = -4.72 \times 10^{-14}$, $SE = 0.02$, $t = 0.00$, $p = 1.00$; lp vs. anom: $\beta = -0.013$, $SE = 0.02$, $t = -0.54$, $p = 0.59$), nor an interaction between coherence and predictability (hp vs anom: $\beta = 2.30 \times 10^{-14}$, $SE = 0.04$, $t = 0.00$, $p = 1.00$; lp vs. anom: $\beta = -0.008$, $SE = 0.04$, $t = -0.20$, $p = 0.84$). The ANOVA showed a significant difference in the goodness-of-fit between the baseline and surprisal models for the pre-critical ROI, with better model fits when the experimental conditions were included ($\chi^2 = 26.50$, $df = 5$, $p < 0.001$).

For the post-critical ROI, I found a significant predictability effect (hp vs. anom: β = –0.86, *SE* = 0.06, *t* = –13.47, *p* < 0.001; lp vs. anom: β = –0.56, *SE* = 0.06, *t* = –8.78, *p* < 0.001). Here, I observe a graded effect of predictability on average surprisal, with the lowest average surprisal for the highly predictable trials, then the less predictable trials and finally the highest average surprisal for the anomalous trials. This aligns with prior evidence on the relationship between predictability and surprisal (N. J. Smith & Levy, 2013; Szewczyk & Federmeier, 2022). There was a non-significant effect of coherence (β = –0.07, *SE* = 0.05, *t* = –1.55, *p* = 0.12), and no interaction was observed between coherence and predictability (hp vs anom: β = 0.03, *SE* = 0.11, *t* = 0.30, *p* = 0.77; lp vs. anom: β = –0.0007, *SE* = 0.11, *t* = –0.006, *p* = 0.99). Comparing the post-critical ROI's baseline and full models, I found a better model fit with the full model (χ^2 = 170.03, *df* = 5, *p* < 0.001). These results highlight that while surprisal is impacted by coherence, the effect is only significant at the pre-critical word ROI, which differs from the previous results found with humans, where the significant effect of coherence emerged at the pre-critical ROI and extended across the sentence to the post-critical ROI while remaining significant.

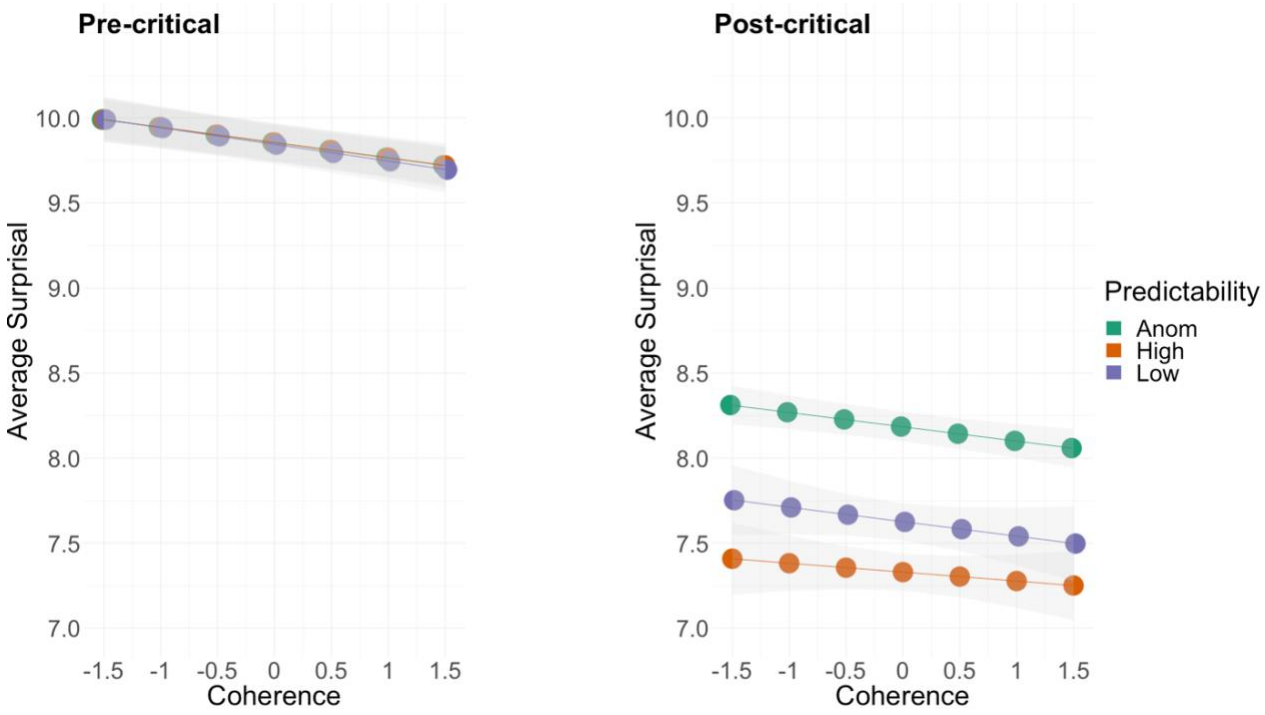

Figure 3.3. Results of predicting average surprisal from coherence and predictability in the pre-critical (left) and post-critical (right) ROIs; shaded areas depict the standard error.

3.4.2. Predicting RTs with surprisal

Experiment 1

For my analyses predicting RTs, I found a significant effect of surprisal in the pre-critical ROI (β = 0.01, *SE* = 0.0003, *t* = 45.90, *p* < 0.001), such that as surprisal increases, the log-transformed RTs also increase. I also found a significant effect of coherence (β = –0.008, *SE* = 0.002, *t* = –4.76, *p* < 0.001), such that as coherence increases, log-

transformed RTs decrease, similar to the human analyses reported in Chapter 2. However, the effect of predictability was not significant (β = 0.001, *SE* = 0.0009, *t* = 1.16, *p* = 0.25), nor was the interaction between coherence and predictability (β = –0.0002, *SE* = 0.002, *t* = –0.15, *p* = 0.88). Additionally, the effects of word length (β = –0.003, *SE* = 0.001, *t* = –2.05, *p* = 0.04) and Zipfian frequency (β = 0.04, *SE* = 0.001, *t* = 30.35, *p* < 0.001) reached significance. The likelihood ratio test results comparing the models with and without surprisal demonstrated a significantly better fit to the data for the model with surprisal ($\chi^2(1)$ = 2088.70, *p* < 0.001).

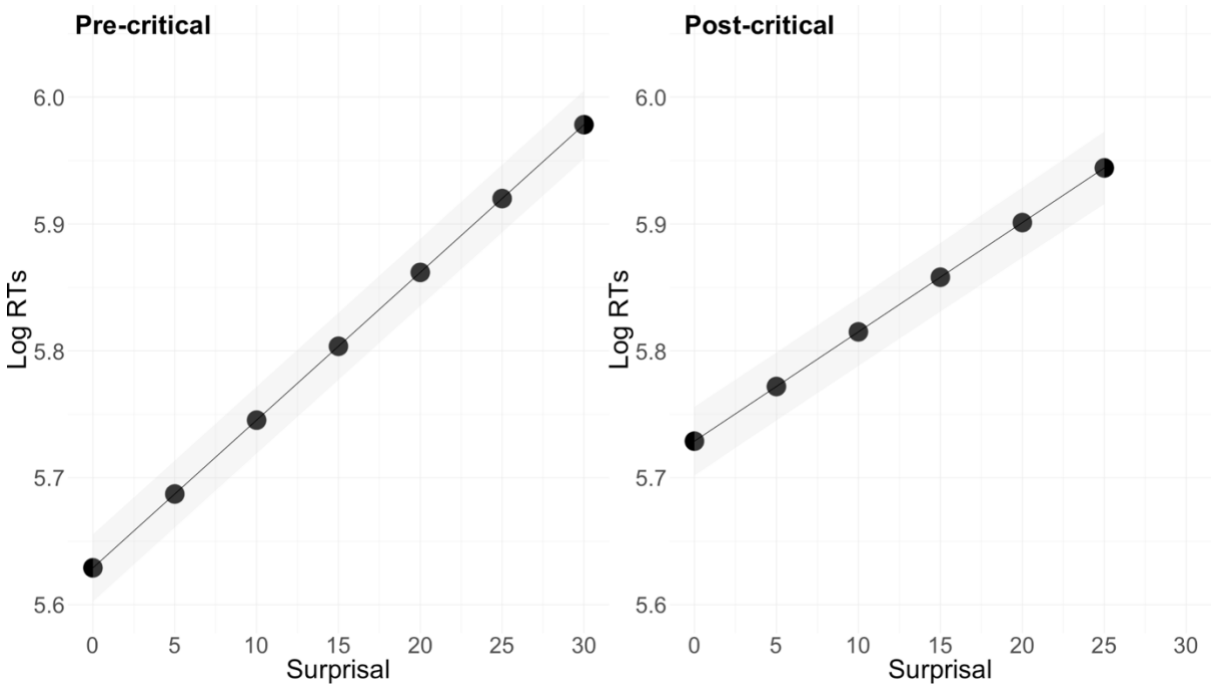

Figure 3.4. Surprisal effects from the results of predicting log RTs from surprisal and the experimental conditions for Experiment 1 in the pre-critical (left) and post-critical (right) ROIs; shaded areas depict the standard error.

For the post-critical ROI, I found a significant effect of surprisal (β = 0.009, *SE* = 0.0006, *t* = 14.91, *p* < 0.001), with longer log-transformed RTs associated with an increase in surprisal. I also found significant effects of coherence (β = –0.009, *SE* = 0.002, *t* = –4.24, *p* < 0.001), and predictability (β = –0.004, *SE* = 0.001, *t* = –3.45, *p* < 0.001), such that greater coherence and predictability trend with shorter RTs. The extended effect of coherence from the pre-critical ROI is similar to the results of Chapter 2, yet the predictability effect was not observed there. However, the interaction between coherence and predictability did not reach significance (β = 0.001, *SE* = 0.002, *t* = 0.59, *p* = 0.55). In addition, the effects of word length (β = 0.04, *SE* = 0.002, *t* = 20.46, *p* < 0.001) and Zipfian frequency (β = 0.02, *SE* = 0.002, *t* = 9.94, *p* < 0.002) were significant. The comparison between the baseline and full models revealed a significantly better model fit with surprisal included ($\chi^2(1)$ = 221.75, *p* < 0.001).

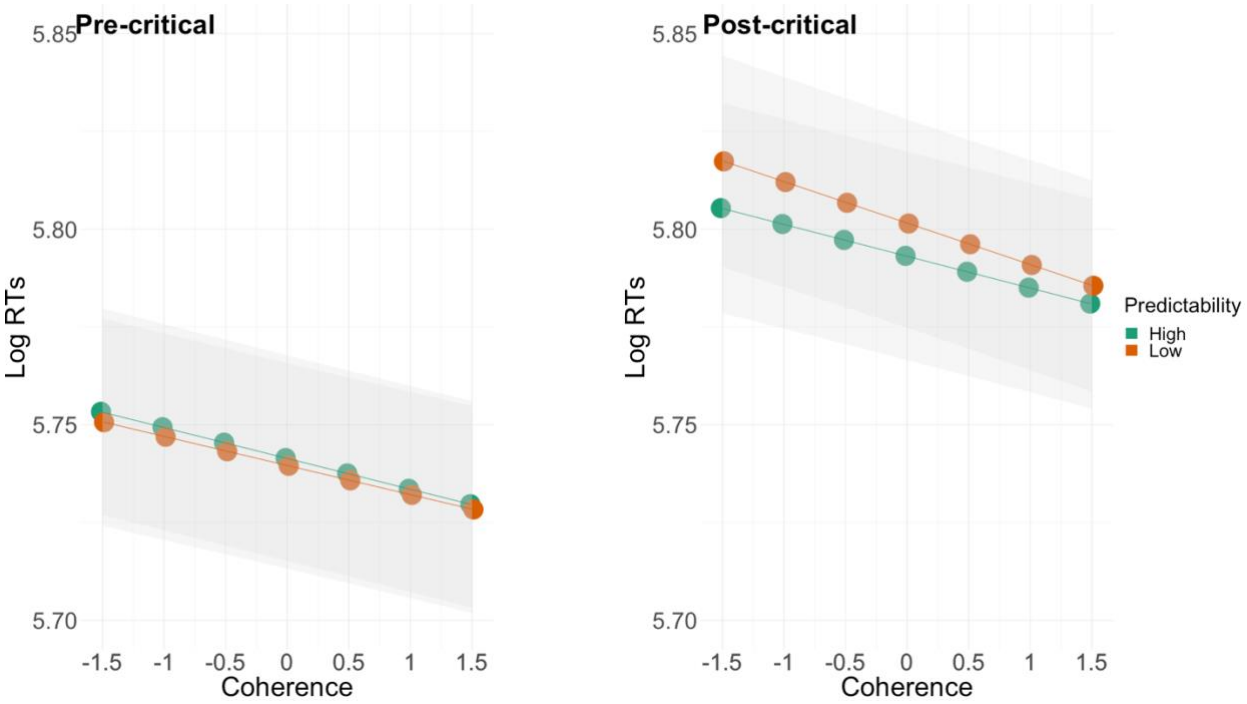

Figure 3.5. Coherence and predictability effects from the results of predicting log RTs from surprisal and the experimental conditions for Experiment 1 in the pre-critical (left) and post-critical (right) ROIs; shaded areas depict the standard error.

Experiment 2

For the experiment 2 analyses, I found a significant effect of surprisal at the pre-critical ROI (β = 0.01, SE = 0.0002, t = 58.61, p < 0.001), demonstrating that as surprisal increases, I observed longer RTs. However, there were no significant effects of coherence (β = –0.002, SE = 0.002, t = –1.30, p = 0.19), predictability (β = –0.0009, SE = 0.0009, t = –1.09, p = 0.27), or their interaction (β = –0.0004, SE = 0.002, t = –0.26, p = 0.79). I did see significant effects of word length (β = –0.005, SE = 0.001, t = –3.61, p < 0.001) and Zipfian frequency (β = 0.05, SE = 0.001, t = 40.22, p < 0.001). Comparing the baseline and full models, I found a significantly better fit to the data for the model including surprisal ($\chi^2(1)$ = 3388.5, p < 0.001).

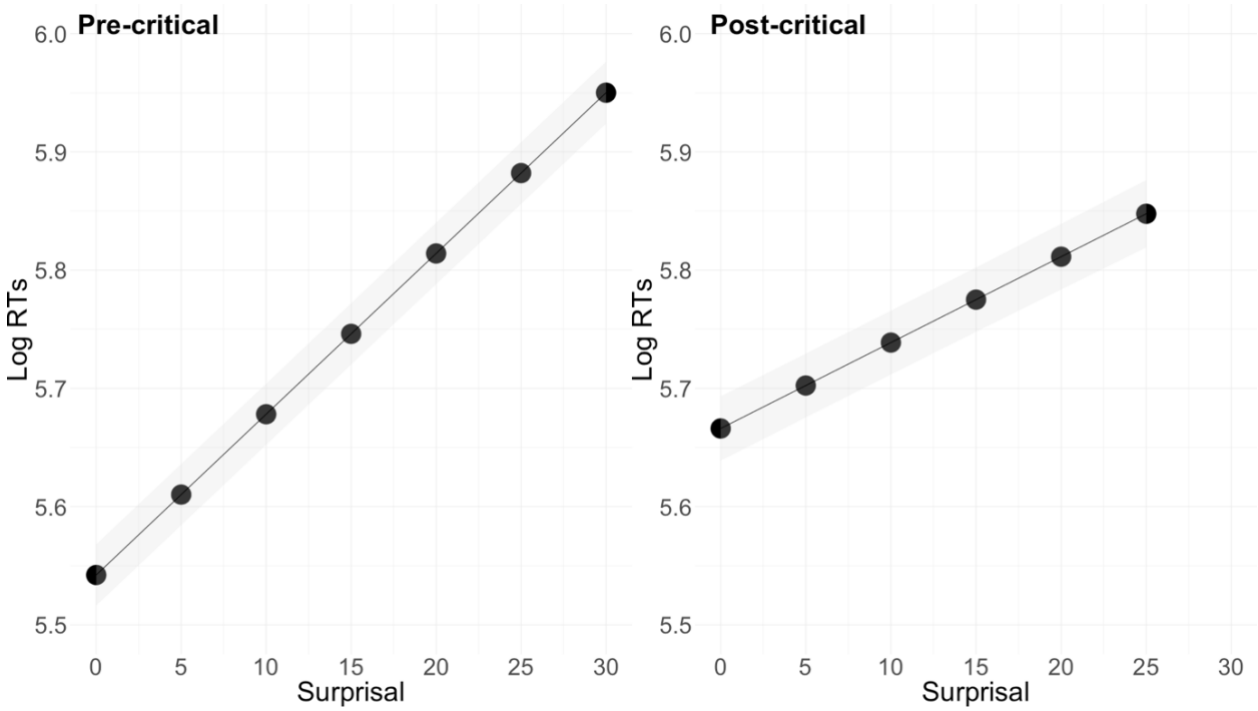

Figure 3.6. Surprisal effects from the results of predicting log RTs from surprisal and the experimental conditions for Experiment 2 in the pre-critical (left) and post-critical (right) ROIs; shaded areas depict the standard error.

For my post-critical ROI, I found significant effects of surprisal (β = 0.007, *SE* = 0.0005, *t* = 13.70, *p* < 0.001), coherence (β = –0.01, *SE* = 0.002, *t* = –5.15, *p* < 0.001) and predictability (β = –0.005, *SE* = 0.001, *t* = –4.58, *p* < 0.001), as well as for word length (β = 0.03, *SE* = 0.002, *t* = 18.37, *p* < 0.001) and Zipfian frequency (β = 0.02, *SE* = 0.002, *t* = 10.01, *p* < 0.001). However, the interaction of coherence and predictability did not reach significance (β = –0.0007, *SE* = 0.002, *t* = –0.34, *p* = 0.74). These results show that as surprisal increases, so do the RTs, whereas, as coherence and predictability increase, RTs decrease. This is in contrast to the results of Chapter 2, where I did not observe significant coherence or predictability effects. For my comparison, I found that the model with surprisal explained additional variance in the data compared to the baseline model ($\chi^2(1)$ = 187.36, *p* < 0.001).

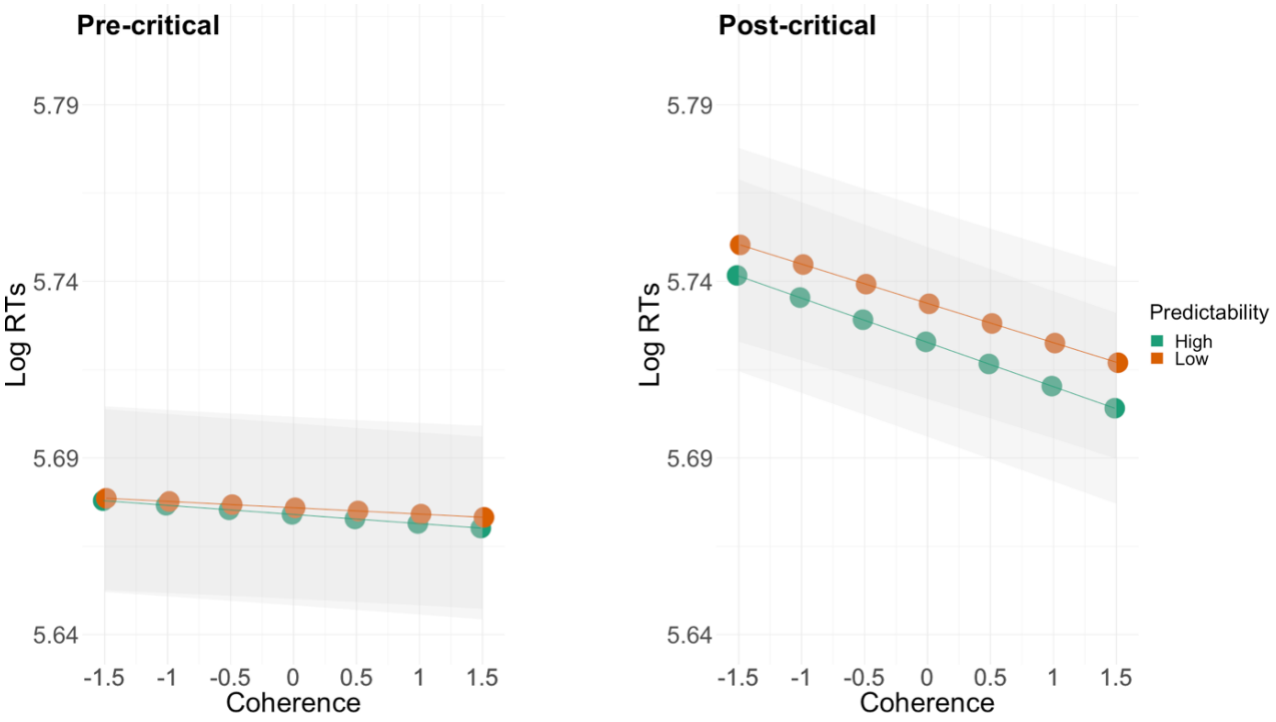

Figure 3.7. Coherence and predictability effects from the results of predicting log RTs from surprisal and the experimental conditions for Experiment 2 in the pre-critical word (left) and post-critical (right) ROIs; shaded areas depict the standard error.

Experiment 3

The pre-critical analyses for Experiment 3 revealed a significant effect of surprisal (β = 0.01, SE = 0.0003, t = 52.99, p < 0.001), with the same trend as previous results. Neither the effects of coherence (β = 0.002, SE = 0.002, t = 0.96, p = 0.34), nor predictability (β = –0.001, SE = 0.0009, t = –1.58, p = 0.12) reached significance. However, there was a significant interaction between them (β = 0.004, SE = 0.002, t = 2.78, p = 0.005). As the experiment materials for the different predictability conditions are the same in this ROI, I believe this to be a spurious result. When comparing the baseline and full models, I observe a better fit to the data with the full model ($\chi^2(1)$ = 2775.8, p < 0.001).

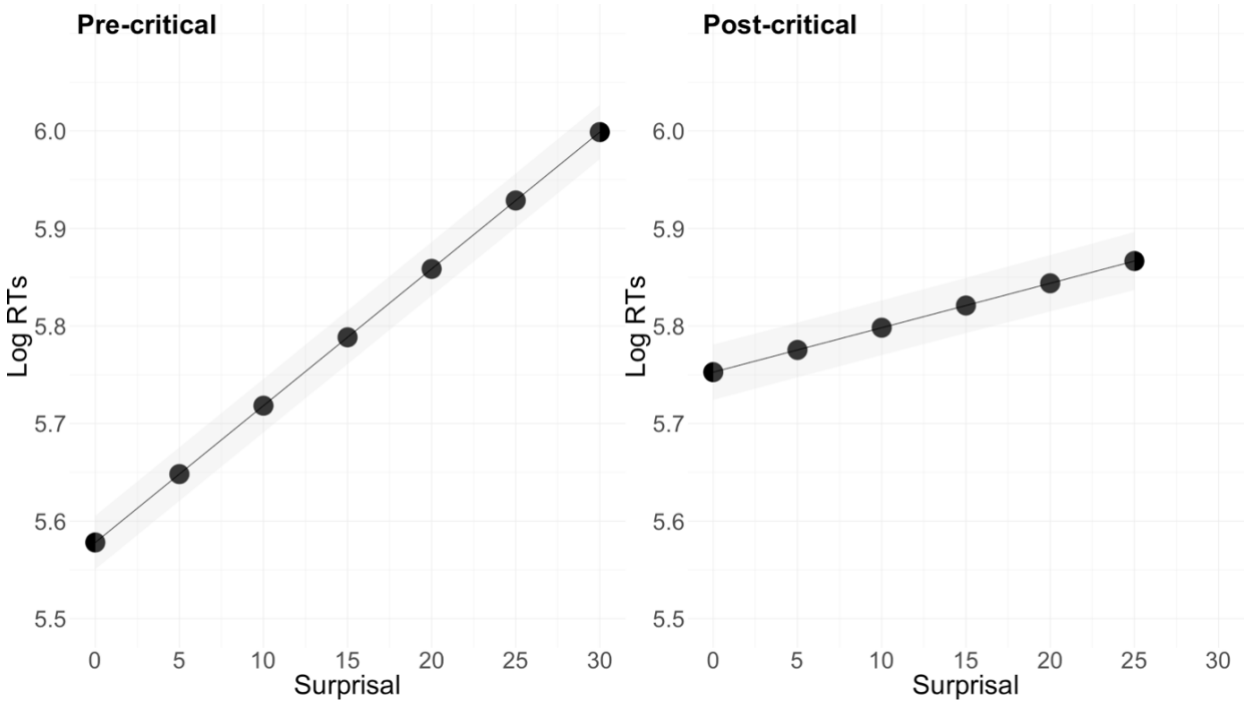

Figure 3.8. Surprisal effects from the results of predicting log RTs from surprisal and the experimental conditions for Experiment 3 in the pre-critical (left) and post-critical (right) ROIs; shaded areas depict the standard error.

The post-critical analyses revealed significant effects of surprisal (β = 0.005, *SE* = 0.0006, *t* = 7.88, *p* < 0.001) and predictability (β = 0.02, *SE* = 0.001, *t* = 12.65, *p* < 0.001). I found that as surprisal increases, the longer the RTs, as well as longer RTs for sentences with anomalous critical words as opposed to highly predictable critical words. There was no significant effect of coherence, however (β = 0.002, *SE* = 0.002, *t* = 0.98, *p* = 0.33). The interaction between coherence and predictability also reached significance (β = 0.007, *SE* = 0.002, *t* = 3.30, *p* < 0.001). These results are the same as those in Chapter 2. I also found significant effects of word length (β = 0.03, *SE* = 0.002, *t* = 16.66, *p* < 0.001) and Zipfian frequency (β = 0.02, *SE* = 0.002, *t* = 10.61, *p* < 0.001). For the comparison between baseline and full models, I found a better fit to the data for the full model that included surprisal ($\chi^2(1)$ = 61.99, *p* < 0.001).

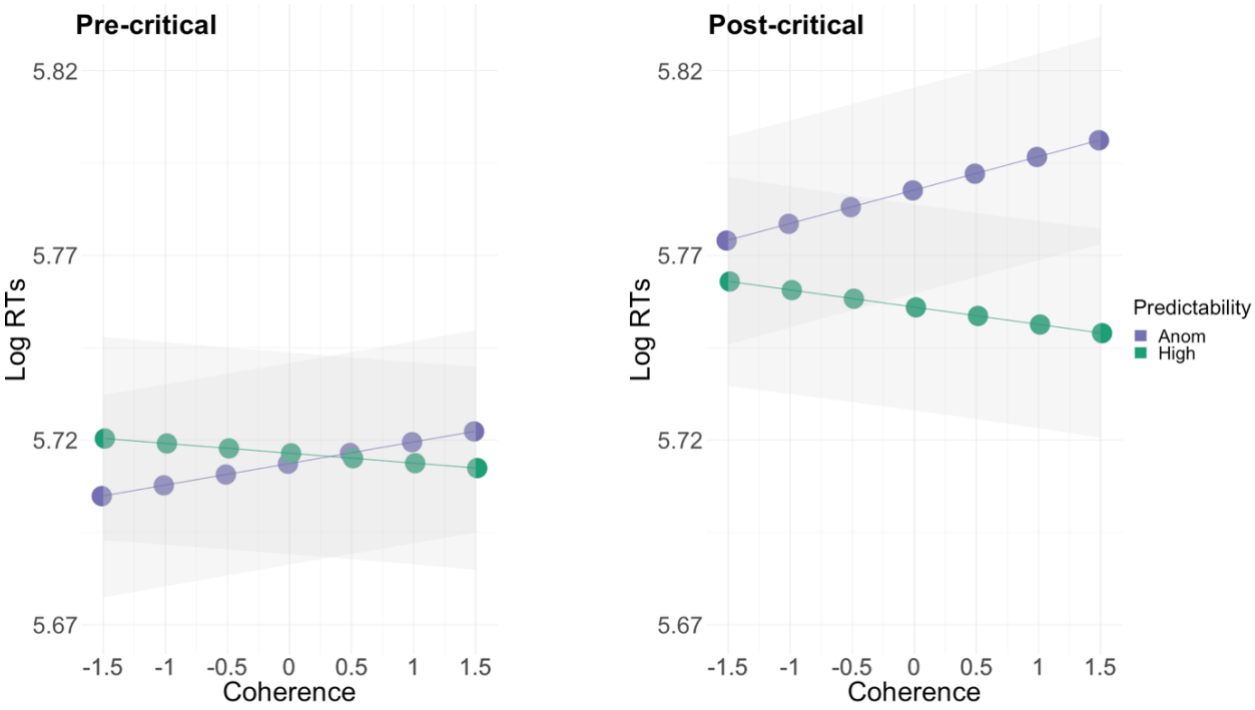

Figure 3.9. Coherence and predictability effects from the results of predicting log RTs from surprisal and the experimental conditions for Experiment 3 in the pre-critical (left) and post-critical (right) ROIs; shaded areas depict the standard error.

3.5. Discussion

In this study, I have investigated how well surprisal values derived from an LLM can account for RTs when people read multi-sentence passages that vary in their levels of internal coherence. I did this by first predicting surprisal from my experimental conditions of coherence and predictability to observe if model predictions are significantly influenced by the same linguistic features that elicited a significant difference in human RTs. I then asked if LM surprisal explains additional variance in predicting my RT data.

The LMEMs predicting surprisal show a coherence effect at the pre-critical ROI, which does not extend into the post-critical ROI. I also observed a graded predictability effect at the post-critical ROI, where I saw the lowest surprisal for trials containing a highly predictable critical word, mid-level surprisal values for the trials with a less predictable critical word, and the highest surprisal for trials containing an anomalous critical word. This finding replicates previous results (Goodkind & Bicknell, 2018; Levy, 2013; Monsalve et al., 2012). Finally, I observed better model fits for all models that included the experimental conditions, suggesting that surprisal is sensitive to aspects of coherence and predictability across multi-sentence stimuli. These results indicate that GPT-2 is sensitive to topic shifts within a passage. From the perspective of language statistics, this suggests that shifts in topics lead to lower predictability.

Meanwhile, the results of the RT LMEMs for Experiment 1 demonstrate significant effects of surprisal in both the pre-critical and post-critical ROIs. I also observe a coherence effect across the ROIs, with a predictability effect emerging at the post-

critical ROI. No interaction between coherence and predictability was found. The coherence effect matches the results from Chapter 2, while the predictability effect is novel. For the model comparison, I find a better model fit to the data with surprisal included. In Experiment 2, I found a significant effect of surprisal across the ROIs, whereas in the post-critical ROI, the effects of coherence and predictability were significant as well. This is in contrast to the results of Chapter 2, where no significant effects were found. Finally, in Experiment 3, I saw a significant effect of surprisal across the ROIs. In the pre-critical ROI, I see a significant interaction between coherence and predictability. However, I believe this to be a spurious result as the critical word has not yet been presented in this region. In the post-critical ROI, I also observed a significant effect of predictability, as well as a significant interaction between coherence and predictability. These results match with what was found in Chapter 2.

Surprisal does explain some additional variance in the RT data, as we consistently observe large effects of surprisal across the target sentence for each experiment. However, the effects of coherence observed in Chapter 2 still remain. This suggests that the benefit of a highly coherent context is not just due to lower linguistic surprisal. It is possible that there are additional levels of processing, for example, the presence of discourse/situation models, which also contributes to the facilitation effect observed (Zwaan, 2016; Zwaan & Radvansky, 1998). It has been suggested that shifts in topic require reconfiguration of the situation model and this additional processing load could lead to a slowdown in comprehension. One difference between the current results and those from Chapter 2 is the presence of predictability effects in the post-critical ROI for the analyses of Experiments 1 and 2. One possibility is that this is due to the difference in pre-processing of the RT data between the two studies.

A critical question relates to whether the observed facilitation effect of a highly coherent context is solely due to the statistics of language, or if there are additional processing mechanisms that humans use when interpreting such stimuli. My current results suggest that there is a further processing advantage beyond mere statistical relations, such as a possible tracking of events and their changes through a discourse-related model. This is in contrast to recent findings that suggest some discourse-level effects from human processing data can be explained solely through language statistics, without making reference to an external situation model (Michaelov et al., 2023a). A possible reason for this difference could relate to differences in the methods used. In my current study, I modelled RT data, whereas Michaelov and colleagues modelled event-related potentials (ERPs). Moreover, the datasets between the two studies are different, with the Nieuwland and van Berkum (2006) dataset including a manipulation of verb–object animacy relations, whereas the current dataset is manipulated on coherence between the context sentences and target sentence. These methodological differences could have led to the divergent conclusions observed. For example, my sample size is much larger, which may have afforded me greater power to tease apart effects of surprisal from discourse coherence.

Another pertinent question relates to the model used to generate surprisal. In the current study, I used GPT-2 to model my data. There is a continuing debate as to the cognitive plausibility of Transformer-based models and their place within cognitive

modelling (Merkx & Frank, 2021; Michaelov, Arnett, et al., 2024). While effective, Transformer-based LLMs are not cognitively plausible models of language processing. This is due to the attention mechanism, which means that the model can process sequential input while attending to all other words in the context directly. As such, it is not clear why surprisal from these models is a good predictor of human language processing data (Merkx & Frank, 2021). I selected GPT-2 as there is prior evidence that suggests it exhibits a better fit to human language processing data than other pre-trained LLMs (Shain et al., 2024), however this is under debate (Michaelov & Bergen, 2022; Oh et al., 2022; Oh & Schuler, 2023).

Limitations of the current work relate to the dataset and model used. For example, investigating the relevance of other linguistic cues, such as implicatures, across different levels of context would help inform current understanding of how comprehenders make use of various cues to inform their expectations about upcoming input. Moreover, it would be insightful to explore a range of models in assessing fit to the data. Possible areas for future research could include other neural architectures, such as RNNs, to add further to the discussion of cognitive plausibility regarding model architectures. A final interesting direction for future research lies in the use of ERPs to evaluate the same linguistic manipulation, i.e., coherence. As ERPs are more temporally accurate than SPR, evaluating N400 responses to these stimuli would help us to understand what specific processes drive the facilitation observed.

3.6. Conclusion

In conclusion, in this study I asked whether surprisal values from GPT-2 are sensitive to manipulations of coherence, in addition to predictability; as well as investigating whether GPT-2 surprisal can predict human's reading times when presented with narratives differing in coherence and predictability. My findings suggest that surprisal is related to subtle topic shifts, and that it can account for some additional variance when predicting reading times. However, it does not fully account for the patterns observed in the RT dataset, as a main effect of coherence in Experiments 1 and 2, and an interaction between coherence and predictability in Experiment 3 still remained. This suggests that the benefit of a highly coherent context is not explained by simply lowering linguistic surprisal, and hints that there are further processing steps leading to the observed facilitation. One possibility is the existence of discourse-related models that comprehenders use to make sense of such narratives. Future work could explore the pattern of responses to these stimuli using ERPs, which would enable a more temporally accurate representation of the mechanisms at play.

Chapter 4

Leveraging context for perceptual prediction using word embeddings

4.1. Abstract

Word embeddings derived from large language corpora have been successfully used in cognitive science and artificial intelligence to represent linguistic meaning. However, there is continued debate as to how well they encode useful information about the perceptual qualities of concepts. This debate is critical to identifying the scope of embodiment in human semantics. If perceptual object properties can be inferred from word embeddings derived from language alone, this suggests that language provides a useful adjunct to direct perceptual experience for acquiring this kind of conceptual knowledge. Previous research has shown mixed performance when embeddings are used to predict perceptual qualities. Here, we tested if we could improve performance by leveraging the ability of Transformer-based language models to represent word meaning in context. To this end, we conducted two experiments. Our first experiment investigated noun representations. We generated decontextualised ("charcoal") and contextualised ("the brightness of charcoal") Word2Vec and BERT embeddings for a large set of concepts and compared their ability to predict human ratings of the concepts' brightness. We repeated this procedure to also probe for the shape of those concepts. In general, we found very good prediction performance for shape, and more modest performance for brightness. The addition of context did not improve perceptual prediction performance. In Experiment 2, we investigated representations of adjective-noun pairs. Perceptual prediction performance was generally found to be good, with the non-additive nature of adjective brightness reflected in the word embeddings. We also found that the addition of context had a limited impact on how well perceptual features could be predicted. We frame these results against current work on the interpretability of language models and debates surrounding embodiment in human conceptual processing.

4.2. Introduction

Our understanding of the world is shaped by the perceptual information we take in through our experiences (Gibbs Jr, 2005; Rogers & Wolmetz, 2016). This perceptual information can be important conceptually—for example, we know that dark chocolate tastes more bitter than white chocolate. The degree to which such perceptual information shapes semantic representations has been a core debate within the

cognitive sciences (Barsalou, 2008; Louwerse, 2011; Pylyshyn, 1980; Rogers & McClelland, 2004). One key question that has emerged from this concerns how much experiential information can be learnt from linguistic content alone (Dove, 2014). Computational models trained on large language corpora provide an important perspective on this debate. These models appear to derive sophisticated semantic representations from linguistic input alone, with no access to perceptual experience. Language models typically generate high-dimensional representations for words and phrases termed "embeddings", which situate concepts in a semantic space. However, a core criticism of these word embeddings is that their dimensions are difficult to interpret and the degree to which they faithfully represent the perceptual aspects of semantics (like the colour of chocolate) remains uncertain (Chersoni et al., 2021; Ettinger, 2020). One way to tackle this issue is to compare word embeddings with human ratings or neuroimaging data to examine the extent to which they mimic human semantics (Abnar et al., 2018; Ettinger, 2020; Hollenstein et al., 2019; Turton et al., 2020, 2021; Utsumi, 2020). In this vein, the present study aims to test how well model-derived embeddings predict human judgements of the salience of specific perceptual properties of objects, and whether additional linguistic context improves these predictions. In so doing, we hope to provide new insights into the degree to which perceptual knowledge can be acquired from language alone.

4.2.1. Conceptual processing: theories and frameworks

Historically, two opposing theoretical perspectives on the nature of semantic representations have been contrasted. The symbolic account of cognition claims that meaning is extracted from language and abstracted into amodal representations (Pylyshyn, 1980). In contrast, the embodied cognition account proposes that human cognition (and by extension, language) is fundamentally grounded in sensorimotor experiences and systems (Bolognesi & Steen, 2018; Gibbs Jr, 2005). Between these extremes, a number of hybrid perspectives envision roles for both language-derived representations and perceptual grounding in supporting semantic processing (Andrews et al., 2014; Barsalou et al., 2008; Louwerse, 2018). Among these accounts, the nature of the representations themselves, as well as the extent to which perceptual information is vital for the formation of these representations, is under debate (Kiefer & Pulvermüller, 2012). Symbolic and embodied perspectives have been supported by an extensive amount of research that has often fallen along different methodological lines, with much of the evidence for symbolic cognition coming from computational models, while the evidence for embodied cognition can be found in studies with human experiments (Andrews et al., 2014; Louwerse, 2018).

There is now widespread evidence from both behavioural and neuroimaging experiments that supports the idea that perceptual representations are often activated during language comprehension (Hauk, 2016; Kiefer & Pulvermüller, 2012; Louwerse, 2018; Meteyard et al., 2012). For example, response latencies from picture verification tasks have demonstrated that comprehenders are sensitive to the orientation of objects in an image, as well as the shape of objects (Stanfield & Zwaan, 2001; Zwaan et al., 2002). Here, response latencies were shorter for pictures of objects that matched in orientation or shape with that suggested by prior context, than for pictures of objects

that did not match. Meanwhile, evidence from neuroimaging has provided additional insights into the between semantic and perceptual information. For example, Simmons and colleagues (2007) demonstrated that the region of cortex strongly linked to colour perception is also active when processing colour terms presented as the properties of objects (e.g., "GRASS-green"). Similar neuroimaging results have also been obtained for other perceptual modalities, such as action (Hauk et al., 2004).

However, much of this research has traditionally treated conceptual representations as static and context-free, without addressing the flexibility that occurs when concepts are used in different contexts (Hoffman et al., 2013; Yee & Thompson-Schill, 2016). Many words have radically different connotations in different situations. The word "bank" should evoke the percept of a large building when used in the context of a city street, but that of a grassy slope when used in the context of a river. Task also influences embodiment: the same words can engage either the visual system or the motor system, depending on which properties are relevant to the participant's current task (van Dam et al., 2012). Findings like these suggest that people flexibly reshape their semantic representations as they encounter different situations (Barsalou, 1983; Jamieson et al., 2022). They also suggest that the degree to which perceptual information is activated depends on the task we are performing and the context in which that word is presented (Barsalou et al., 2008).

To understand the degree to which perceptual information is embedded in language, many researchers have investigated the capabilities of computational language models that are solely exposed to language input. We review these findings in the next section. For now, it is important to highlight that much of this research has also been conducted from a context-free perspective. Pioneering models like latent semantic analysis (LSA; Landauer & Dumais, 1997), as well as more recent models such as word2vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) and GloVe (Pennington et al., 2014) have been invaluable in discovering the semantic structure present in language. But these models represent each word as a context-independent, static embedding. Given the importance of context in shaping human semantic representation, these context-free representations likely underestimate the semantic information present in language. More recent developments in natural language processing, in the form of Transformer-based models, allow contextualised word embeddings to be generated (Devlin et al., 2018; Misra et al., 2020; Ontanon et al., 2022). These embeddings provide a contextualised representation of a word, such that the embeddings for a polysemous word, such as "bank", will be different across sentences that make use of the distinct senses of the word. In this study, we investigate the types of perceptual information present in different forms of contextualised embeddings.

4.2.2. Interpreting the semantic content of word embeddings

Word embeddings, as vector-based representations of language, are strongly related to the distributional hypothesis. This is the notion that the semantic similarity between two linguistic expressions can be understood as a function of the similarity of the

linguistic contexts that they appear in (Firth, 1957; Harris, 1954; Lenci, 2008). This has widely influenced the cognitive science of semantics, where the distributional hypothesis is also proposed as a cognitive hypothesis for the organisation of meaning. As such, models derived from distributional language data have been used to model multiple aspects of human language processing, such as word association, semantic deficits and categorisation (Bullinaria & Levy, 2007; Griffiths et al., 2007; Vigliocco et al., 2004). Early examples of these distributional models include Latent Semantic Analysis and the Topic model (Griffiths et al., 2007; Landauer & Dumais, 1997), while more recent examples include Word2Vec and GloVe (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; Pennington et al., 2014).

Word embeddings extracted from these models can be represented geometrically in a semantic space, where words that are more semantically related cluster together (Riordan & Jones, 2011). Because of this, semantic similarity has commonly been used as an evaluation metric for distributional representations (Günther et al., 2016; Jones et al., 2006; Lenci, 2008; Lowe & McDonald, 2000). For example, Grand and colleagues used semantic projection to compare the internal representations of word embeddings against human judgements of object categories and properties. The authors constructed scales denoting a semantic feature of interest, say size, and were able to compare the internal representations of word embeddings by creating a featural subspace where size influences the similarity patterns between the embeddings. They found that these feature-wise similarities predicted human feature judgements, concluding that the geometric representation of word embeddings contains rich conceptual knowledge (Grand et al., 2022).

Another approach has focused on learning a mapping between word embeddings and property norms as a way to ground them in interpretable representations (Chersoni et al., 2021; Derby et al., 2019; Făgărăşan et al., 2015; Utsumi, 2020). Previous research using this approach has produced mixed findings on the degree to which word embeddings mimic perceptual aspects of human semantics. Abdou and colleagues (2021) explored BERT, RoBERTa and ELECTRA embeddings for colour words (e.g., “yellow”) using representational similarity analysis (RSA) and linear regression. They found that the embeddings of colour words aligned with the structure of a 3D colour space, CIELAB, concluding that an approximation of the perceptual colour space can be extracted from text alone. However, this success in extracting colour knowledge from embeddings has not been replicated when probing for colour information about objects (e.g., bananas are yellow). Vecchi and colleagues (2017) explored the factors influencing semantic acceptability of novel adjective–noun phrases using a human-generated ratings dataset and compositional distributional semantics. They find that the cosine distance between the composed phrase and its component noun is best able to predict the human acceptability ratings of semantically deviant adjective–noun phrases, such as ‘academic bladder’. Sommerauer and Fokkens (2018) used Word2Vec embeddings to classify objects according to whether they possessed particular features (e.g., is yellow, is dangerous). While functional and behaviour-relevant features were generally classified well, performance was poor for perceptual features, including colour. In a similar vein, Lucy and Gauthier (2017) evaluated word embeddings on how well they predicted perceptual and conceptual features of

concrete concepts, using semantic norm datasets collected from humans as a gold-standard. They tested different types of word embeddings (GloVe and Word2Vec) using the McRae and CSLB semantic norm datasets (Devereux et al., 2014; McRae et al., 2005) and found that the embeddings failed to encode many salient perceptual features of the concepts, in comparison to strictly non-perceptual categories (such as taxonomic and functional features).

The previously mentioned studies used embeddings to predict the presence or absence of binary features (e.g., is yellow vs. is not yellow). Meanwhile, other studies have tried to predict continuous ratings of the importance or relevance of different types of information. For example, Chersoni and colleagues (2021) trained a neural network to learn the mapping from word embeddings (both count-based models e.g., PPMI, GloVe and prediction-based models e.g., SGNS, BERT) to vectors devised from human ratings. The human-based vectors were taken from the Binder dataset (Binder et al., 2016), which contains ratings on the relevance of 65 semantic features to 535 concepts. Participants were asked to rate the relevance of each semantic feature for a particular concept. The 65 features were selected to represent core modalities of information processing from the neuroimaging literature. For example, the dataset includes features focusing on specific sensory and motor experiences (e.g., the relevance of shape and motion), as well as affective experiences (e.g., happy and sad). The models were tested on their ability to predict values over all 65 features for unseen words. The authors found that social, causal and cognition features were generally better predicted than sensorimotor features, consistent with the idea that language is a critical source of information for these types of semantic features (Borghi et al., 2019). Within the perceptual domain, some somatosensorial features were well-predicted (such as colour and shape), whereas others were less well captured (such as bright and dark) (Chersoni et al., 2021).

A number of other studies have used the Binder dataset to investigate the knowledge represented in word embeddings. Turton and colleagues (2020, 2021) used both Word2Vec and BERT embeddings to predict the feature rating vectors in the Binder dataset, finding that some perceptual features were again well predicted (such as colour and shape), but others were less well represented (such as bright, dark and slow). They also demonstrated that the mappings learnt between word embeddings and the feature rating vectors can be extrapolated to a wider vocabulary than the original dataset, whilst keeping the semantic relationships between features intact. Utsumi (2020) conducted a similar experiment evaluating the mappings between the feature rating vectors from the Binder dataset and word embeddings using three types of distributional embeddings (SGNS, GloVe and PPMI) that were derived from training on two different corpora (COCA and Wikipedia). Similar to Chersoni et al. (2021), they found that social, causal and cognition features were better predicted, with perceptual features less likely to be represented in word embeddings. For example, features relating to the brightness or speed of a concept were predicted poorly. However, Utsumi concluded that some ability to predict perceptual information was present for domains such as shape, vision and sound. Utsumi (2020) also investigated the prediction performance separately among concrete and abstract concepts. Predictions were poorer for abstract words across all feature types, except for emotion features. Perhaps

unsurprisingly, perceptual features were predicted particularly poorly for abstract words, consistent with the long-standing view that abstract concepts have few perceptual associations (Paivio, 1990). Taken together, these studies provide some evidence that embeddings are able to predict some types of perceptual information associated with concepts, though prediction of purely perceptual features is poorer than for other types of semantic information.

Why do perceptual qualities seem to be less well-represented in word embeddings? In the previous section, we noted that the engagement of perceptual processing during language comprehension is highly dependent on context. Distributional methods for obtaining word embeddings (e.g., Word2Vec, GloVe) have been decontextualised (also known as "static"). This means that the entire word representation is encoded as a single vector, abstracted across all the different contexts in which the word is used. As such, static embeddings capture the most significant semantic properties and relationships that are most reliably represented across contexts. This means that less salient information, such as perceptual qualities, may not be well-represented. However, further advances in machine learning have led to Transformer-based LLMs, where a word's embedding changes depending on the linguistic context in which it is presented (Vaswani et al., 2017). These models have more sophisticated embeddings with the potential to encode context-specific information, such as different word senses and contextually relevant properties. Researchers have begun to test the predictive abilities of contextualised word embeddings using the Transformer-based BERT model (Devlin et al., 2018). Turton et al. (2021) generated BERT embeddings for concepts by sampling 250 sentences containing each word in the Binder dataset. These sentences were input into BERT, providing 250 different contextualised word embeddings for each target word. A single, context-free representation was then obtained for each word by averaging over its 250 embeddings (for similar approaches, see Bommasani et al., 2020; Chersoni et al., 2021; Vulić et al., 2020). Turton et al. (2021) found that BERT embeddings created in this way outperformed static embeddings in predicting the feature rating vectors from Binder et al. (2016), suggesting that simply aggregating embeddings over many contexts leads to representations that better capture human experience with concepts. They went on to demonstrate that handpicking 10 sentences which matched the sense of the word as used in the Binder dataset improved this result further.

The benefit of contextualised embeddings was also demonstrated when assessing the effects of specific contexts. Here, Turton and colleagues (2021) used a dataset of semantic feature ratings for property-object pairs (e.g., "abrasive lava"; "abrasive sandpaper"; Van Dantzig et al., 2011) to investigate how the presence of a specific context influences prediction performance. In the original study, participants were asked to provide ratings on five separate scales representing each perceptual modality in answer to the prompt: "To what extent do you experience [object] being [property]" (Van Dantzig et al., 2011). Turton and colleagues first fed property-object pairs into the Transformer models and extracted the word embeddings for the property words in the context of a specific object. They then compared the performance of these embeddings against an averaged property embedding (i.e., averaged across two object contexts) and a static baseline Numberbatch embedding in predicting feature ratings

for the property-object pairs (Speer et al., 2017). For example, they compared how well the extracted embeddings for "abrasive" predicted the perceptual feature ratings for "abrasive lava" and "abrasive sandpaper". They found that the contextualised Transformer embeddings outperformed the mean Transformer embeddings and the Numberbatch embeddings. This study provided a first indication that contextualising embeddings with a specific use case can lead to more effective prediction of perceptual properties. In the present study, we build on this idea in two experiments. In the first, we investigate whether the prediction of human-generated ratings on the perceptual properties of nouns is improved when contextualised embeddings are generated using a context that specifically primes for the desired perceptual feature (e.g., "the brightness of charcoal", "the shape of charcoal"). We investigated both ratings on the relevance of the feature to the noun (either brightness or the relevance of shape), and ratings on the perceived brightness of the noun. In the second experiment, we extend this to adjective-noun phrases (e.g., "dark charcoal"), which have rarely been studied. Here, we attempt to predict ratings of the perceived brightness of conceptual combinations. We tested whether contextualised embeddings better predict the perceptual ratings of such phrases and whether language models compose the meanings of adjective-noun phrases in a similar way to humans.

4.3. Experiment 1

In Experiment 1, we compared contextualised and decontextualised word embedding performance on how well they can predict human-generated ratings of the perceptual qualities of nouns. We explored the prediction performance of word embeddings from a Distributional Semantic Model (Word2Vec) and a Large Language Model (BERT). We investigated these issues using two specific perceptual features as test cases: brightness and shape. We chose brightness as it has not been predicted well in previous research (Chersoni et al., 2021; Utsumi, 2020), and therefore represents a challenging test case for examining the extent of perceptual feature representation in embeddings. This is in contrast to shape, which has been well-predicted (Chersoni et al., 2021; Turton et al., 2021; Utsumi, 2020). We also selected brightness because it allowed us to make use of a critical dataset that contained ratings for both unmodified nouns, and adjective-noun combinations (the Solomon and Thompson-Schill dataset), which allowed us to explore the nature of conceptual combinations in context.

First, we used bright and dark ratings from the Binder dataset, which has been most commonly used in previous research on this topic (Chersoni et al., 2021; Turton et al., 2020, 2021; Utsumi, 2020). Binder et al. (2016) obtained ratings for many different features so this dataset contains a large number of items for which brightness is not a salient (e.g., "chair") or relevant feature (e.g., "advantage"). Second, we used the brightness dataset of Solomon and Thompson-Schill (2020). S&T-S only explored brightness and therefore they collected ratings for a smaller set of concepts for which brightness/darkness was a relevant and salient feature (e.g., "diamond", "grey", "charcoal"). By comparing perceptual prediction across these two datasets, we were able to test the degree to which findings from the Binder dataset generalise to other datasets, which are tailored to the specific feature under investigation. For comparison,

we also predicted ratings on the relevance of shape to the concepts. We selected this perceptual feature as it has previously been reported to be well predicted by BERT embeddings (Chersoni et al., 2021; Turton et al., 2020). We used the Binder ratings for this investigation.

Figure 4.1 shows an overview of our experiment pipelines. For the Word2Vec embeddings, we extracted embedding representations of the nouns from Google's pre-trained model that was trained on a section of the Google News dataset (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). We then used these embeddings as input to a feed-forward neural network to predict the perceptual feature ratings for that noun. Word2Vec generates a single representation of each word and therefore these embeddings are decontextualised and static. In contrast, we used BERT's capacity for contextualisation when extracting our BERT embeddings. Here, we make use of Google's pretrained models, extracting embeddings from both the BERT base model and the BERT large model, which differ in dimensionality (Devlin et al., 2018). Following previous studies, we included a context-free condition, which contained an averaged BERT embedding from multiple sentence contexts (Bommasani et al., 2020; Chersoni et al., 2021; Turton et al., 2021; Vulić et al., 2020). This context-free condition can also be thought of as a prototype because it derives an abstract representation of each word across many different instances (Hampton, 2015; Rosch & Mervis, 1975). To do this, we identified 250 sentences that contained the noun of interest from the one Billion Words Benchmark corpus (Chelba et al., 2014). We then extracted the BERT embedding for the noun in each sentence and averaged these. We also had a contextually prompted condition, where we included a contextual prompt targeted towards the perceptual feature of interest. For our brightness investigations, we made use of two different prompts, "the brightness of…" and "the colour of…"; while we only used one prompt for the shape investigations: "the shape of…". We separately presented these prompts to BERT and extracted the embedding for the noun. We expected the contextually prompted BERT embedding to better predict the perceptual feature of interest, compared to both the Word2Vec and context-free BERT embeddings, as it would bias the embedding towards feature-relevant aspects of the semantic representation.

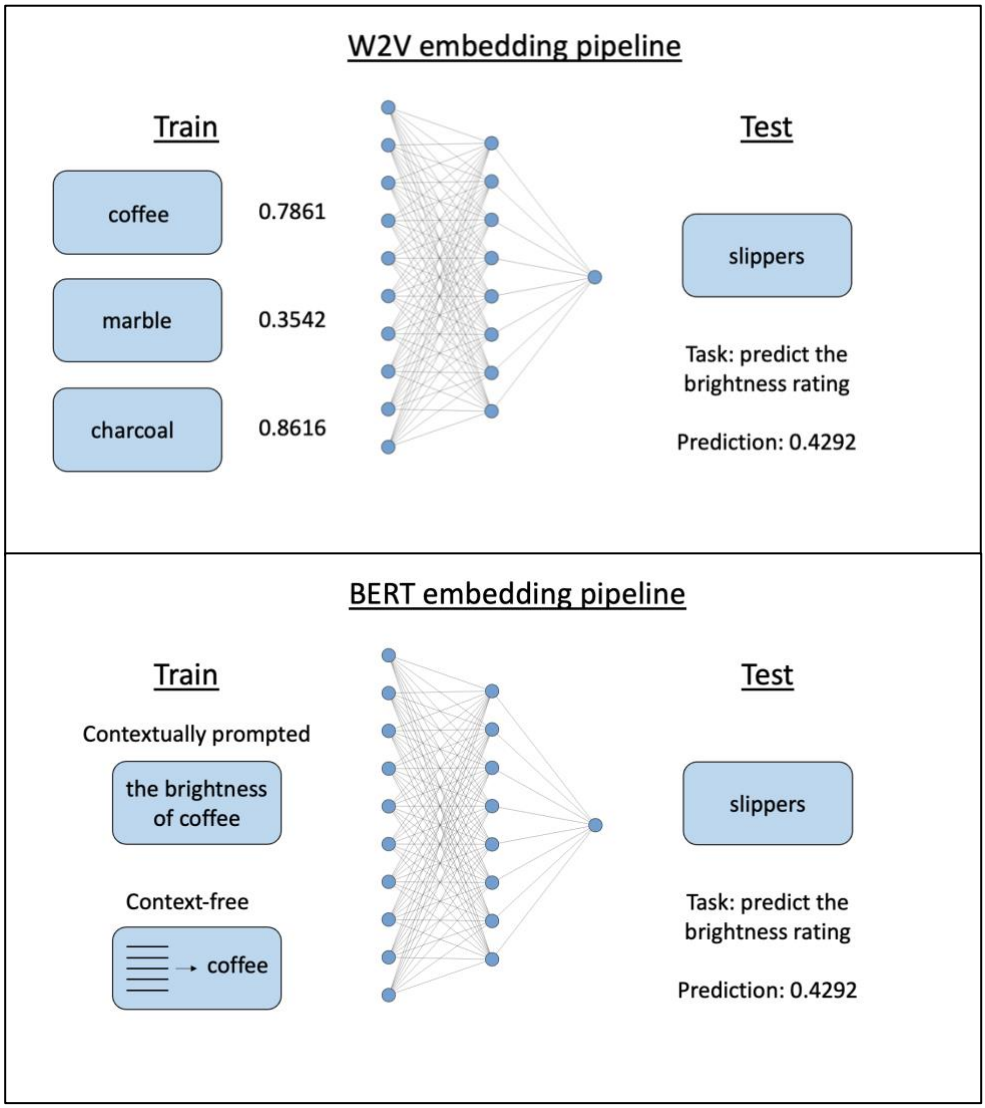

Figure 4.1. Experiment 1 pipelines.

4.3.1. Methods

All associated code and data can be accessed here: https://osf.io/ca4wm/?view_only=6d2f88c1f0b347f1b2be72a22b5b9e7e

Datasets

We evaluated model performance on predicting human ratings of perceptual qualities from two datasets. In the **Solomon and Thompson-Schill (S&T-S) dataset**[3], Solomon and Thompson-Schill (2020) asked participants to rate the brightness of 45 nouns on a scale of 0-50 (brightest to darkest), both with and without modifiers. Participants (n=100) made these judgments by moving a slider, with a bar showing a greyscale colour spectrum running from 0 (white) to 50 (black). We scaled the ratings to lie between 0-1 (brightest to darkest). The nouns in this dataset were specifically chosen to represent the entire spectrum of brightness values (e.g., "charcoal" vs. "snow"), and to include

[3] The S&T-S dataset can be accessed here: https://osf.io/7uwn9/

concepts for which brightness was a relevant feature. We removed 3 adjectives from the original dataset to focus solely on nouns ("black", "white" and "grey"). The authors also collected ratings for adjective-modified versions for each noun, which we use in Experiment 2.

The **Binder dataset**[4] consists of semantic ratings collected by Binder et al. (2016). Here, the authors aimed to create a dataset of conceptual feature ratings that was informed by known modalities of neural information processing. The authors settled on 65 semantic features. Participants (n=1743) were asked to rate the relevance of each semantic feature to a word's meaning on a scale from 1-6. Each participant was assigned a single word and provided the ratings for all 65 features. As this task was crowdsourced on Amazon Mechanical Turk, the authors included a quality metric to ensure that participants focused on the task. As such, the correlation between a participant's vector and the group average vector of a word was calculated and if this did not exceed a minimum value of r = .5, the participant was discarded. The original Binder dataset included two features related to the brightness of a concept: ratings on the degree to which each concept is visually bright or dark. To transform these features into a brightness metric similar to that used in the S&T-S dataset, we first scaled each dimension between 0-1, subtracted the scaled dark dimension from the scaled bright dimension, and then transformed the output to ensure that all values fell between 0 and 1. This way, the spectrum of brightness ratings mimicked Solomon and Thompson-Schill's, such that stereotypically bright words (e.g., "sun") had a low rating (0.133), while stereotypically dark words (e.g., "crow") had high ratings (0.949). It is important to note that Solomon and Thompson-Schill (2020) specifically selected items for which brightness is a highly relevant property. In contrast, the Binder dataset includes many items which are not strongly associated with a particular level of brightness; hence many items were clustered around the midpoint of the brightness spectrum. For our second perceptual feature, the relevance of shape, we also used ratings from the Binder dataset.

We used two versions of the Binder dataset, one which contained only concrete nouns and another that contained both abstract and concrete nouns. The concrete version of the dataset was created by filtering the original dataset on type and super category. We filtered type to only contain items classed as a "thing" or "event", with super category filtered to include "artifacts", "living objects", "natural objects" and "physical states". This resulted in a dataset of 274 concrete nouns. For the investigations that included abstract items, we additionally filtered the super category by "abstract entity", "event", and "mental entity". This resulted in a dataset of 433 items, with 275 concrete items and 159 abstract items. We chose to explore the performance of concrete concepts alone as Utsumi (2020) found poor performance for the prediction of perceptual properties for abstract concepts. Therefore, we wanted to test the effect of excluding these.

Embeddings

[4] The Binder dataset can be accessed here: https://www.neuro.mcw.edu/index.php/resources/brain-based-semantic-representations/

We used three sets of pre-trained word embeddings: Word2Vec, BERTbase and BERTLarge. This allowed us to make a comparison between word embeddings that can leverage contextual information (contextualised) versus those that are static and independent of context (decontextualised). Moreover, we specifically used pre-trained embeddings, rather than training our own, as we wanted to assess the predictive ability of publicly available embeddings that researchers may use to shed light on the conceptual representations of words (Günther et al., 2019; Pereira et al., 2016). We also included **one-hot** vectors as a baseline comparison for each experiment in order to test the influence of our feedforward neural network training. Here, each word is encoded as a vector whose dimensionality is equal to the number of words in the vocabulary. For a single word representation, the node associated with that word (e.g., "charcoal") will be "on", while all other nodes are switched "off". As such, one-hot vectors represent a simple lexical-coding scheme that contains no semantic information.

Word2Vec embeddings are an example of static word embeddings, meaning that they assign a fixed vector to each word, regardless of the context in which the word is being used (Mikolov, Chen, et al., 2013). We used pre-trained embeddings from a Word2Vec model trained on a section of the Google News dataset (~100 billion words). The dimensionality of the embeddings is d=300, and previous evidence demonstrated that they are a good fit to human semantic judgements across a range of tasks (Pereira et al., 2016). Further details on these pre-trained embeddings can be found on the Google Code Archive (https://code.google.com/archive/p/word2vec/).

BERT embeddings are an example of contextualised word embeddings and were extracted from BERT (Bidirectional Encoder Representations from Transformers), a large language model based on the Transformer architecture (Devlin et al., 2018). The model is trained on a masked language modelling task, where language samples are provided with 15% of the words masked at random and the model is trained to predict the masked words. We used the Hugging Face Transformers API to access two different pre-trained models, which differ in size (bert-base-uncased and bert-large-uncased) (Wolf et al., 2020). These models were pre-trained on two corpora (~3 billion words), BookCorpus and English Wikipedia, and tokenized using WordPiece, a subword-based tokenization algorithm. We extracted the corresponding word embedding from the last hidden layer, and if the noun was split into separate sub-words, we extracted the sub-word embeddings and averaged them to represent the entire word (dimensionalities: **BERTbase**: d=768; **BERTLarge**: d=1024). We chose to extract embeddings from the last hidden layer as previous research has demonstrated that semantic features are better represented by higher layers (Jawahar et al., 2019; Turton et al., 2021).

To create our **context-free** condition, we replicated the method from Turton et al. (2021) for creating a "static" version of BERT embeddings. For each concept, we randomly extracted ~250 sentences containing the target word from the One Billion Words Benchmark corpus (Chelba et al., 2014) accessed via HuggingFace (https://huggingface.co/datasets/lm1b). To do this, we started with a target word (e.g., "coffee") and the shuffled train partition of the corpus (n_sentences=30,301,028). We then searched through the corpus, using string match to identify sentences that

contained our target word, and saved examples of the first 250 sentences that were found. These sentences were then cleaned to remove extraneous punctuation marks and whitespace. For each concept, we ran each of the 250 tokenized corpus sentences through BERT, located the position of the target word in the sentence and extracted its word-level embedding (or the averaged subword-level embedding). We then averaged the 250 embeddings together, which was used as input to our neural network.

To create our **contextually prompted** condition, we generated specific prompts for the nouns depending on the feature to be predicted. For brightness, we initially used "the brightness of [noun]". However, this phrasing could be considered somewhat unusual and unnatural for some concepts in the dataset (e.g., most people would describe crows as being dark in colour, not dark in brightness). As such, we also tested a second prompt, "the colour of [noun]", which is twice as frequent in the Google n-grams corpus[5]. We present results from both prompts and used the best-performing prompt in comparisons with other embeddings. For predicting shape, we used "the shape of [noun]". We then ran these tokenized phrases through BERT, located the target word at the end of the phrase, and extracted its word-level embedding (or averaged subword-level embedding) as input to our neural network. As such, we had three versions of the contextually prompted condition: brightness, colour and shape.

Model

Following a similar approach to other studies (Sommerauer & Fokkens, 2018; Turton et al., 2020, 2021; Utsumi, 2020), we trained a three-layer feed-forward neural network to predict human feature ratings from our word embeddings. The model was implemented in PyTorch (Paszke et al., 2019) and consisted of an input layer (dimensions dependent on the input embedding), a ReLU activation function, a dropout layer with p=0.2, a hidden layer (dimensions dependent on the investigation type) and a single output unit (d=1) with a sigmoid activation function to normalise predictions between 0 and 1. For our training procedure, we used k-fold cross validation where k=10. In each fold, models were trained with 90% of concepts and were then tested on their ability to predict the relevant feature for the remaining 10% of concepts. The hyperparameters used for model training were: learning rate = 0.01, bias = -2, momentum = 0.9 and a weight decay = 10^{-6}. We optimised using stochastic gradient descent. We also performed hyperparameter tuning for the number of hidden units and number of epochs to train, using gridsearch and nested cross-validation (k=3). We kept the tuned hyperparameters the same across specific experimental comparisons to ensure fairness.

Evaluation

To reduce random noise, for each investigation, 10 different models were initialised with randomised starting weights. Each of the 10 models was trained and tested according to the 10-fold cross-validation scheme described earlier. Differences in model performance across the 10 implementations were small (standard deviations of errors

[5]https://books.google.com/ngrams/graph?content=the+colour+of%2C+the+brightness+of&year_start=2000&year_end=2019&corpus=en-2019&smoothing=3&case_insensitive=true

across models are presented in the Appendix). We obtained a single prediction for each concept by averaging the predictions of the 10 models. We evaluated performance of the different embeddings using mean squared error (MSE) and R^2. The MSE was calculated as the average of squared errors between model prediction and the human rating for each concept. The R^2 for the correlation between model predictions and the human ratings was calculated by fitting an ordinary least squares regression model. Additionally, we ran statistical tests on comparisons of interest for both brightness and shape, outlined below. To evaluate these comparisons, we used the Wilcoxon signed-rank test, which is a paired-samples, non-parametric test, comparing squared errors for each noun. For shape and brightness, we compared:

- Word2Vec vs context-free BERT
- Context-free BERT vs contextually prompted BERT

For the brightness experiments, we also compared:

- "Brightness" contextually prompted BERT vs "colour" contextually prompted BERT

For experiments using the Binder dataset, we also compared performance for:

- Brightness vs shape

4.3.2. Results

Solomon and Thompson-Schill Dataset

Figure 4.2 presents an overview of the noun results for each embedding set on the S&T-S dataset, along with the MSE and R^2.

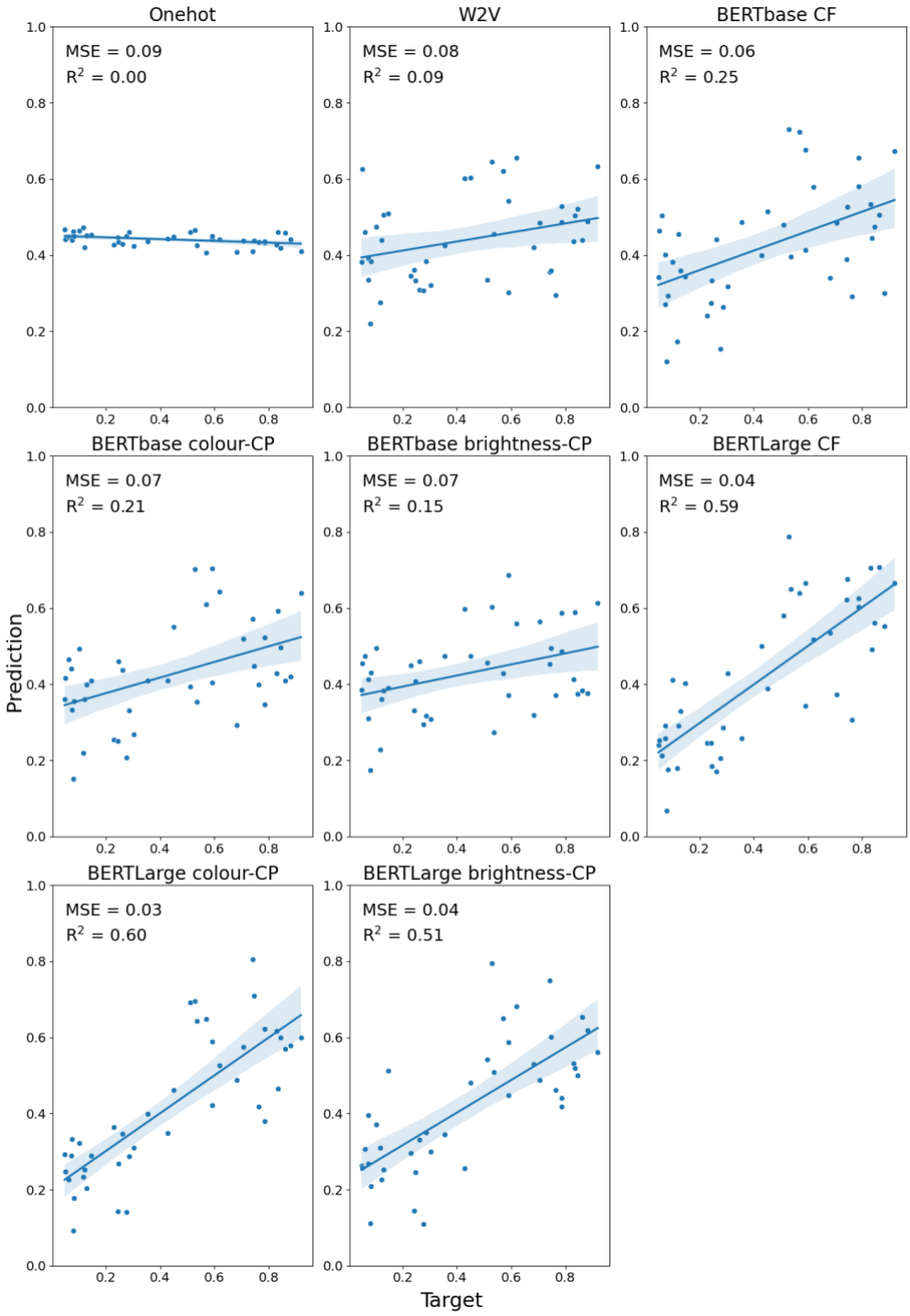

Figure 4.2. Predicted vs target brightness values for the S&T-S nouns. Shaded areas indicate 95% confidence interval for regression. CF= context-free, CP= contextually prompted.

As expected, the model trained with one-hot vectors which contain no semantic information was unable to predict brightness of the target concepts. Predicted brightness values in this model were tightly clustered around 0.5 and were unrelated to target concept brightness. Models trained with embeddings performed better. In general, BERTLarge embeddings appeared to have the best performance, with R^2 values ranging from 0.5-0.6. Thus, although previous studies have reported poor prediction of

brightness/darkness, here we found that embeddings can predict these to a reasonable extent. This may be because we are focusing particularly on a set of nouns for which brightness is a highly relevant feature. For the decontextualised embeddings, we found that the context-free BERTLarge embeddings performed significantly better than Word2Vec ($p=6.97 \times 10^{-6}$). This was also true for the context-free BERTbase embeddings, but to a lesser extent ($p=0.04$). Turning to the contextually prompted embeddings, we found no significant difference between the type of prompt ("the colour of" vs. "the brightness of") for either BERTbase ($p=0.20$) or BERTLarge ($p=0.37$). As such, we chose to compare the "colour" contextual prompt for our comparison with the context-free embeddings as the R^2 was higher. We found no significant difference between the contextually prompted and context-free conditions for either BERTbase ($p=0.33$) or BERTLarge ($p=0.66$). This suggests the addition of a contextual prompt does not improve performance for predicting brightness on the S&T-S dataset.

Binder Dataset: Brightness

Figure 4.3 (concrete nouns only) and Figure 4.4 (concrete and abstract nouns) present the MSE and R^2 for each embedding when the model was trained to predict brightness for the Binder dataset. In general, both dataset configurations produced a similar pattern of results. Overall, the best performing embedding was the context-free BERT condition with both concrete and abstract nouns (MSE= 0.01, R^2= 0.25).

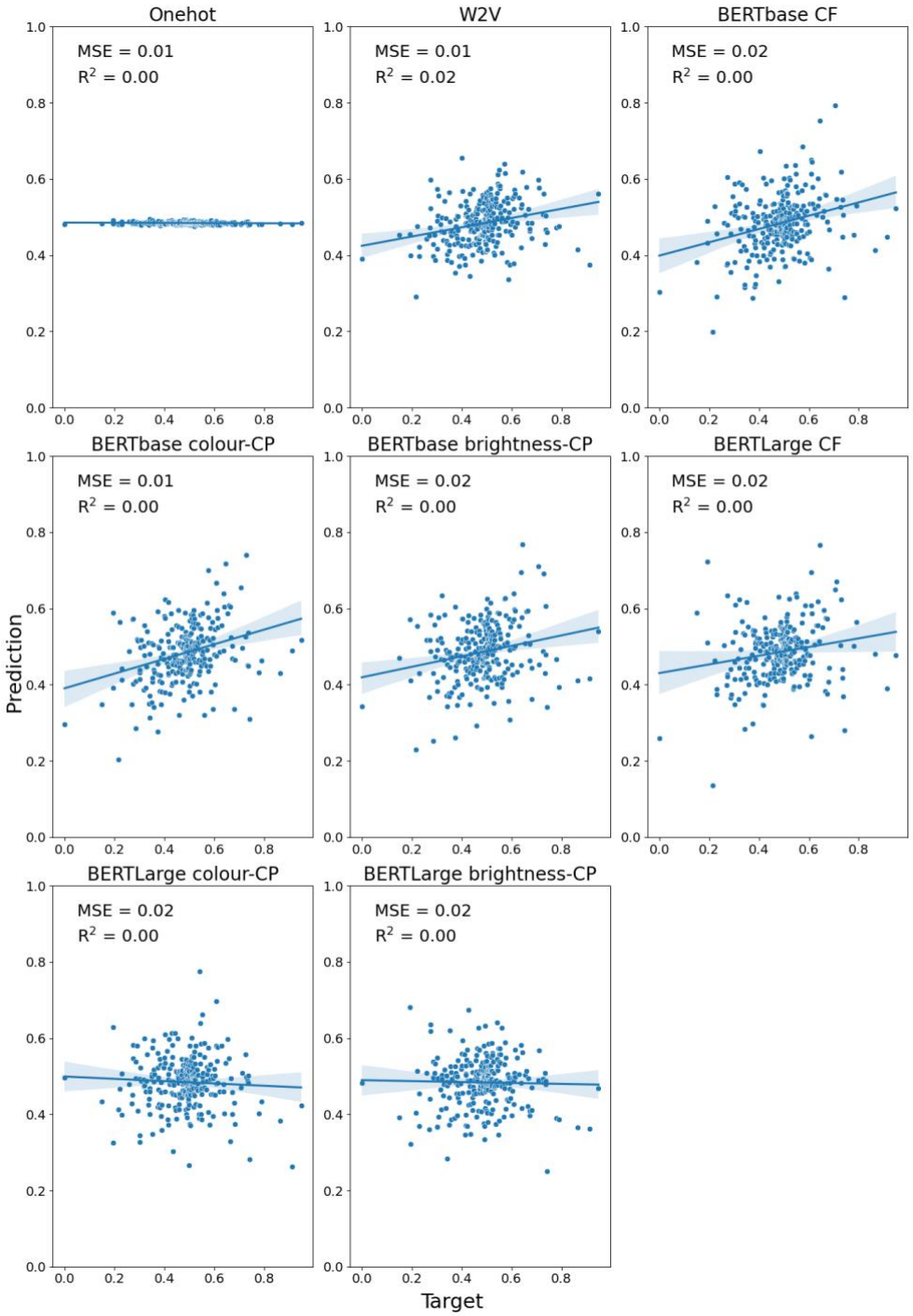

Figure 4.3. Predicted vs target brightness values for the Binder concrete-only nouns. Shaded areas indicate 95% confidence interval for regression. CF= context-free, CP= contextually prompted.

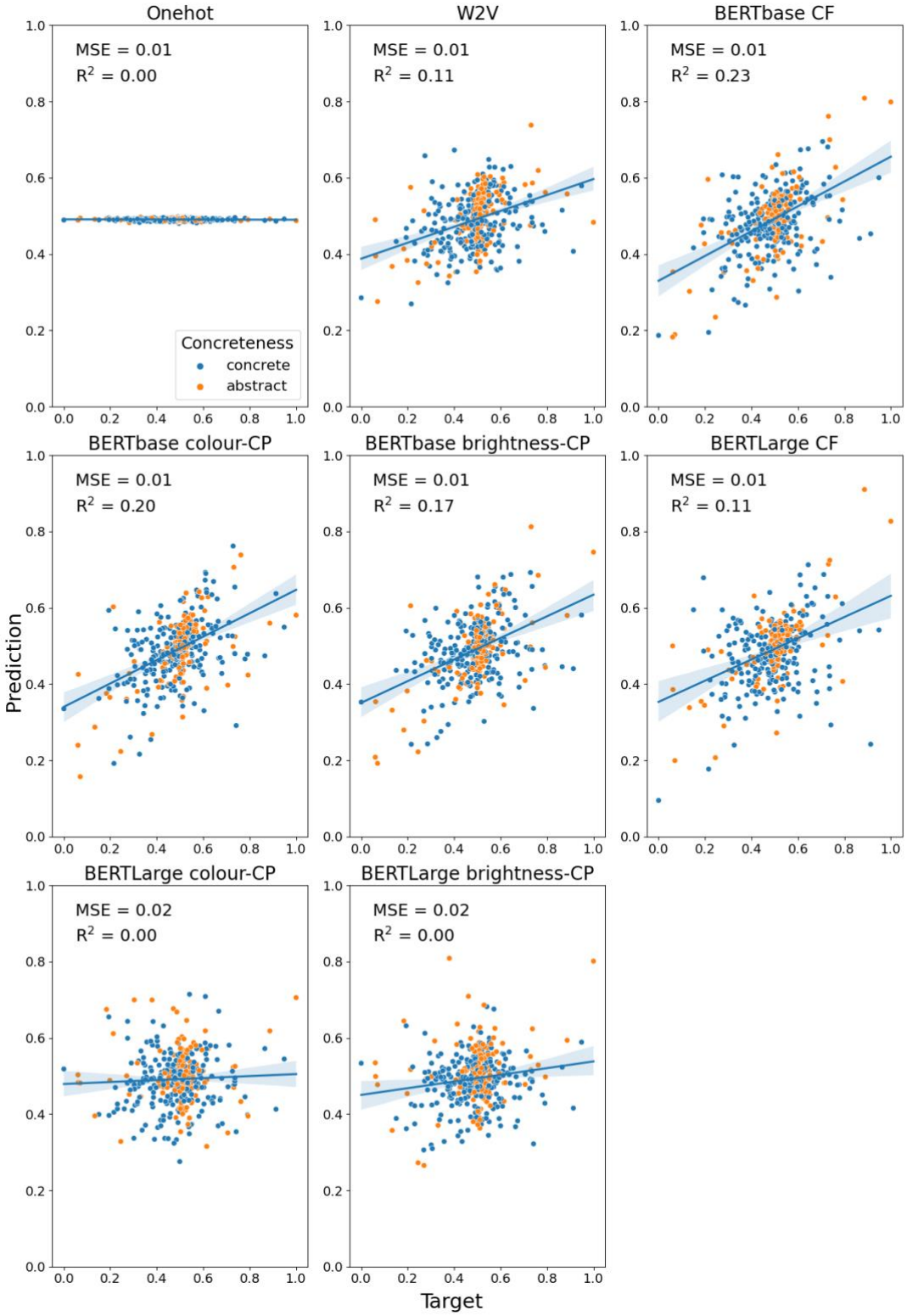

Figure 4.4. Predicted vs target brightness values for the Binder concrete (blue) and abstract (orange) nouns. Shaded areas indicate 95% confidence interval for regression. CF= context-free, CP= contextually prompted.

Overall, we found poorer performance across the embedding types when predicting brightness in the Binder dataset compared with the S-T&S dataset (maximum R^2 = 0.25 vs. 0.63). This suggests that brightness information is not as well represented

in the embeddings for this larger dataset that contains many items where brightness is not a highly relevant property. In our comparison of the decontextualised embeddings, we found that the context-free BERT embeddings outperformed Word2Vec on the concrete+abstract dataset for both BERTbase (p= 0.005) and BERTLarge embeddings (p= 0.05). However, this was not the case for the concrete-only dataset (base: p= 0.33; large: p= 0.34). This suggests that the improved predictive ability of the context-free BERT embeddings over Word2Vec may stem from better prediction of abstract concepts in this dataset. For the comparison of the prompt in our contextually prompted conditions, we found that embeddings with the "colour" contextual prompt performed better than embeddings with the "brightness" contextual prompt for BERT embeddings in most cases, though this was only significant for the BERTbase concrete-only dataset (p= 0.01). As such, we used the "colour" prompted embeddings for our statistical comparison with the context-free embeddings. We found a statistical difference between the context-free and contextually prompted conditions for BERTLarge embeddings (concrete: p= 0.02; concrete+abstract: p= 2.48×10^{-7}), but contrary to expectations, the context-free embeddings had better prediction performance. Moreover, for the BERTbase versions, we found no statistical difference in prediction performance (concrete: p= 0.09; concrete+abstract: p= 0.96).

Binder Dataset: Shape

Next, we move onto the comparison for predicting the relevance of shape for different concepts. See Figure 4.5 (concrete-only) and Figure 4.6 (concrete+abstract) for an overview of the MSE and R^2 for this investigation.

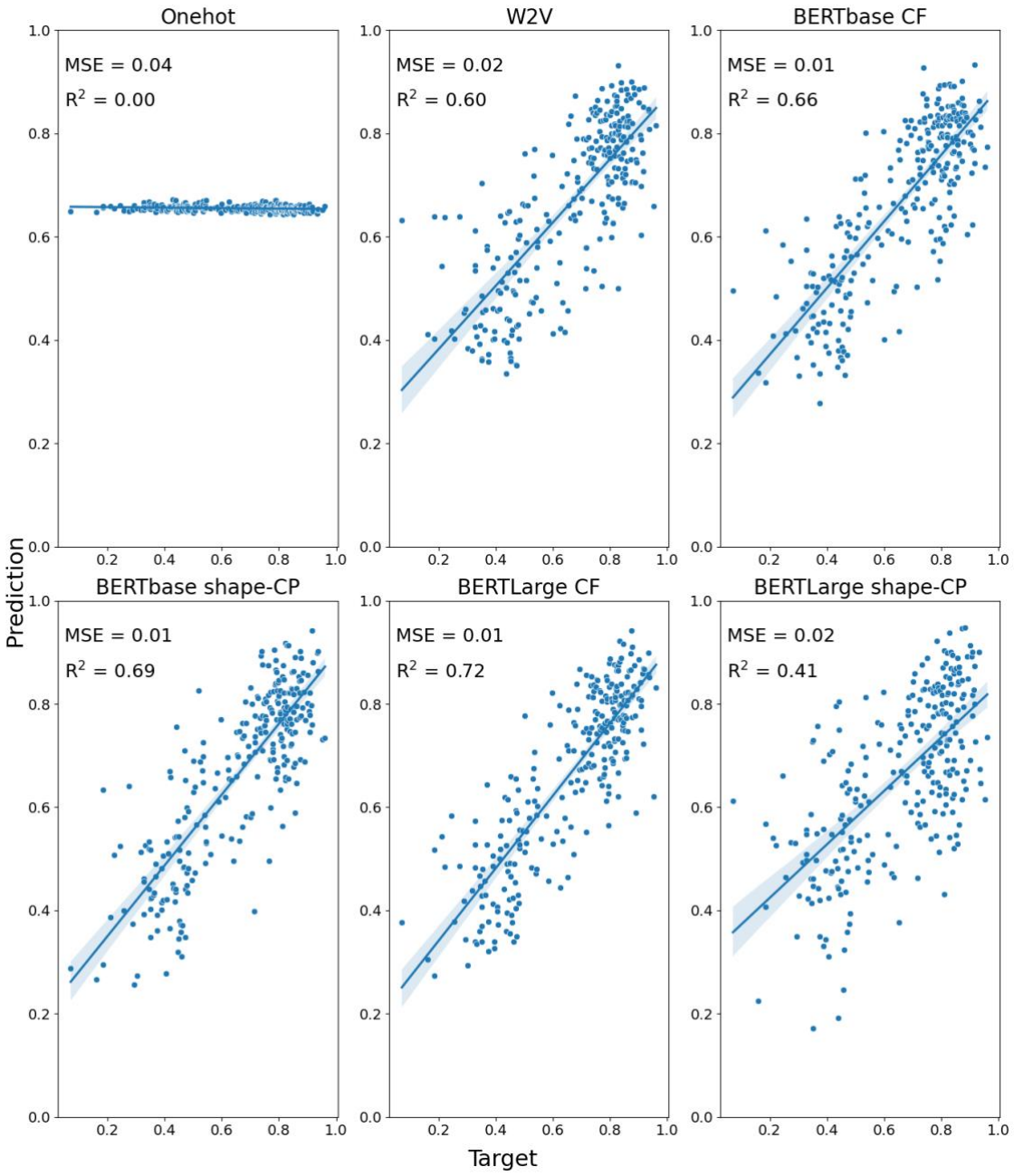

Figure 4.5. Predicted vs target shape values for the Binder concrete-only nouns. Shaded areas indicate 95% confidence interval for regression. CF= context-free, CP= contextually prompted.

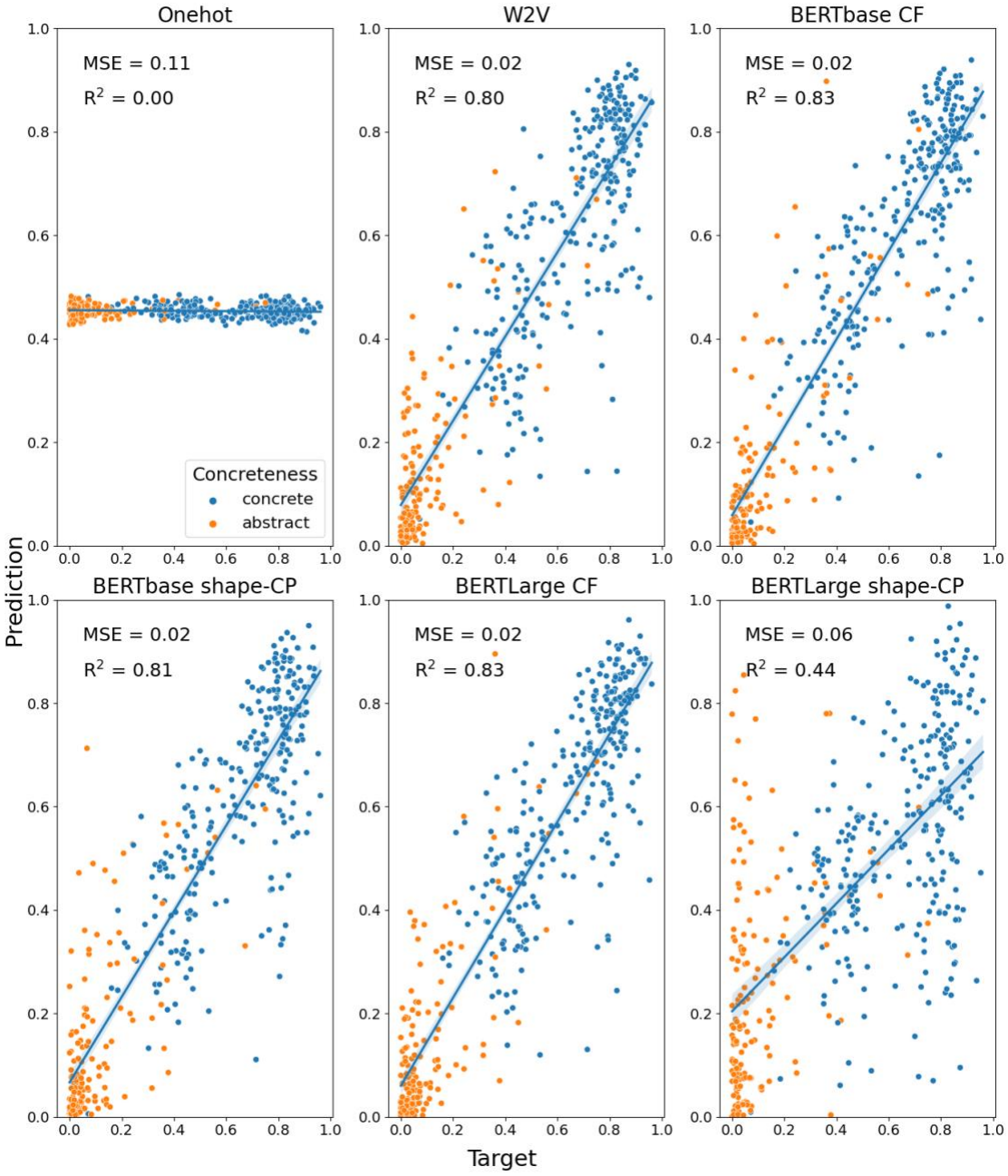

Figure 4.6. Predicted vs target shape values for the Binder concrete (blue) and abstract (orange) nouns. Shaded areas indicate 95% confidence interval for regression. CF= context-free, CP= contextually prompted.

As expected based on previous studies, embeddings showed very high performance in predicting the relevance of shape (Turton et al., 2021; Utsumi, 2020). The best performing embeddings were the context-free BERT embeddings on the concrete+abstract dataset, with no difference in performance between the base and large versions (base: MSE= 0.02, R^2= 0.83; large: MSE= 0.02, R^2= 0.83). In general, the addition of abstract nouns improved model performance for most embedding types. This is likely related to the fact that shape values for abstract concepts cluster at the lower end of the spectrum. This strong distinction between concrete and abstract nouns may have helped bootstrap learning of the mapping between nouns and the relevance of shape.

Looking at the decontextualised embeddings, we found significant differences between Word2Vec and BERT embeddings for the majority of conditions, with the context-free BERT embeddings performing slightly better than the Word2Vec embeddings (base: concrete+abstract: p= 0.004; large: concrete: p= 0.002; concrete+abstract: p= 0.02). This was not the case for the BERTbase embeddings on the concrete-only dataset (p= 0.23). Moreover, in our comparison between context-free and contextually prompted BERT embeddings, we found that the context-free BERTLarge embeddings had significantly greater prediction performance than the contextually prompted embeddings (concrete: $p = 9.64 \times 10^{-9}$; concrete+abstract: $p = 4.33 \times 10^{-18}$). However, this trend was not apparent for the BERTbase embeddings (concrete: p=0.21; concrete+abstract: p=0.06). As such, it appears that the addition of a feature-specific prompt can lead to worse prediction performance for the perceptual feature of shape.

Finally, we also statistically compared the best-performing embeddings for brightness and shape features on both dataset configurations. For the concrete-only dataset, we selected the context-free BERTbase embedding for brightness and the context-free BERTLarge embedding for shape. Here, there was no significant difference in performance (p= 0.63). For the concrete and abstract dataset, we chose the context-free BERTbase embeddings for both features. Here, we found a significant difference between performance, with better prediction performance for shape than for brightness (p= 0.01). This suggests that shape is a better represented perceptual feature in these types of word embeddings than brightness, which is consistent with previous results (Turton et al., 2020, 2021).

4.3.3. Discussion

In Experiment 1, we found good prediction of brightness for the S&T-S dataset, but not for the Binder dataset, suggesting that even difficult perceptual properties can be predicted when focusing on a subset of nouns where the feature is particularly salient. Replicating previous findings, we also found very good prediction performance for the relevance of shape to concepts (Turton et al., 2021; Utsumi, 2020). In general, we found that context-free BERT embeddings outperformed Word2Vec embeddings, which aligns with prior findings that aggregated contextualised representations are better predictors of semantic features than static embeddings (Bommasani et al., 2020; Turton et al., 2021). However, contrary to our predictions, we found the contextually prompted embeddings often performed worse than the context-free embeddings. We consider reasons for this in the General Discussion. In addition, onehot vector representations were entirely unable to predict any perceptual ratings, unlike the pre-trained word embeddings. This confirms that success in prediction is a consequence of the semantic information present in the word embeddings and not the neural network we used to map from embeddings to ratings.

4.4. Experiment 2

Until now, contextualised embeddings have primarily been tested on their ability to predict the perceptual properties of nouns. However, how these embeddings represent

the perceptual properties of multi-word expressions, such as adjective-noun phrases, is an important consideration. Linguistic theory states that adjectives modulate the properties of nouns and they frequently do so in a non-uniform manner (Solt, 2019). In particular, the linguistic literature makes a distinction between subsective adjectives whose meaning is context-sensitive, and therefore depends on the comparison class that they modify such as “tall”, and intersective adjectives that have a more context-insensitive meaning, such as colour terms (Demonte, 2019; Partee, 2007). As a subtype of subjective adjectives, relative gradable adjectives such as “slow”/ “tall” are additionally characterised by vagueness and have even been theorised to not directly denote properties. Instead, it has been argued that the denotation is only ascribed meaning during the process of composition (C. Kennedy, 2007, 2012).

Traditional theories on the composition of conceptual combinations include the selective modification model, which specifically focuses on adjective–noun combination. Here, it is assumed that concepts are represented as schema-like structures with sets of dimensions and corresponding values, similar to prototype theory (Hampton, 2015; Rosch & Mervis, 1975; Rumelhart, 1980). During combination, an adjective’s primary feature is reweighted onto the noun concept (E. E. Smith et al., 1988; E. E. Smith & Osherson, 1984). However, criticisms of the selective modification model note that the process of combination is more complex, especially when expanded to other conceptual combinations such as noun–noun compounds (Hampton, 2015; Murphy, 1988). In contrast, the concept specialisation view states that combination occurs through specialisation of the head noun concept when one of its “slots” is filled by the modifying concept (Cohen & Murphy, 1984; Murphy, 1988, 2004). The theory emphasises the role of general background knowledge and reasoning in forming conceptual combinations (Murphy, 2004). In summary, theories on conceptual combination illustrate the notion that the combinatorial process itself is highly idiosyncratic and dependent on the composing concepts (Coutanche et al., 2019). This suggests that adjective–noun phrases may be a particular case in which static embeddings are insufficient in capturing the underlying semantics, since the process of combination shifts the representation of both words in an unpredictable fashion.

Adjective–noun phrases are a valuable way to isolate the integration process of conceptual combination as they are independent of additional processes of property selection. Solomon and Thompson-Schill (2020) demonstrated this by asking people to rate the perceptual feature of brightness for adjective-noun pairs. They found that the adjectives “dark” and “light” modified people’s perceptions of the brightness of nouns, but they did not do so in an additive fashion: the adjectives had more of an impact on some nouns than others. For example, for a mid-brightness noun, such as “paint”, there was a big difference between the perceived brightness of its light (“light paint” = 0.112) and dark (“dark paint” = 0.867) versions. For other nouns with a more extreme and invariant perception of brightness, such as “charcoal”, the adjectives had less of an effect (“light charcoal” = 0.565; “dark charcoal” = 0.930). Examples such as these would be a clear case where we would expect contextualised embeddings to have an advantage in perceptual prediction, compared to static embeddings, in capturing these complex interactions between adjectives and nouns. As such, for our second experiment, we explored whether contextualised embeddings could accurately predict

properties for modified adjective-noun pairs, and whether targeted prompts towards the relevant property improves this behaviour.

To test this, we again used Word2Vec embeddings, contextually prompted BERT embeddings and context-free BERT embeddings. For the Word2Vec embeddings, we extracted each of the adjective and noun embeddings and concatenated them to represent adjective–noun phrases. As we wished to evaluate predictions for the unmodified noun, alongside the "light" and "dark" versions, we paired the nouns with an adjective that was uninformative with regards to brightness. We chose "heavy" for this because it is a high-frequency adjective (similar to "light" and "dark") that can be applied to objects, whilst conveying no information about their brightness. Our training paradigm for this investigation tested each model's prediction performance on unseen adjective–noun phrases. For the BERT embeddings, we again compared contextually prompted embeddings with context-free embeddings. As we were interested in multi-word expressions, we used BERT's class (CLS) token to represent the entire phrase (see Methods). For the context-free conditions, we extracted the CLS token for the adjective-noun phrase alone, while for the contextually prompted conditions, we used the CLS token for the feature-prompted phrase (e.g., "the brightness/colour of dark paint"). Figure 4.7 presents an overview of our experiment pipelines for Experiment 2.

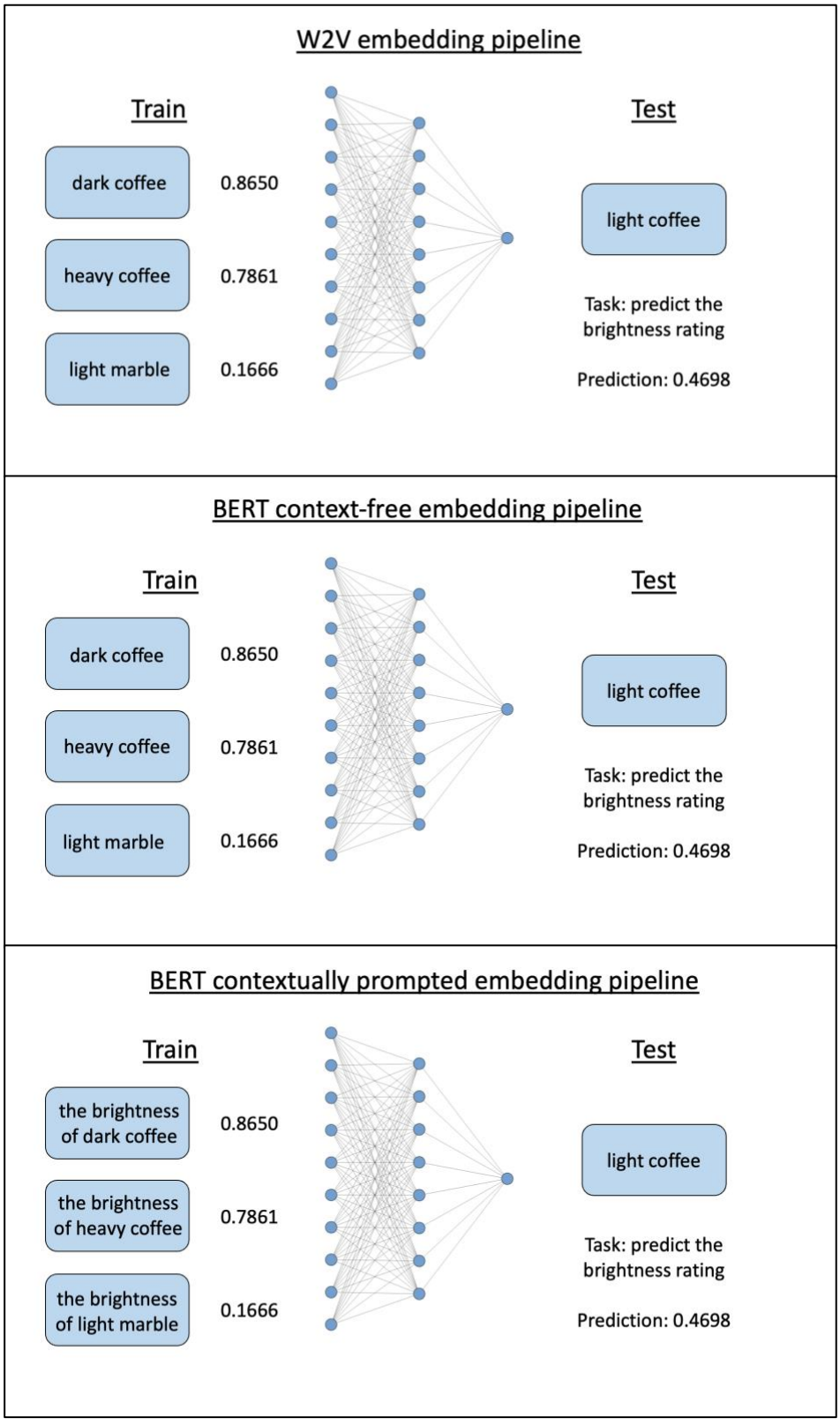

Figure 4.7. Experiment 2 experiment pipelines.

4.4.1. Methods

Dataset

We used the S&T-S dataset for our adjective-noun investigations. For this study, we used all three rating types: the unmodified noun ratings (as used in Experiment 1), and the "dark" and "light" adjective-modified ratings. These ratings were originally scored between 0-50 (bright to dark), which we transformed to between 0-1 (bright to dark). As there are 42 nouns in the dataset, there were a total of 126 adjective-noun phrases.

Embeddings

For the **Word2Vec** embeddings, we used the pre-trained embeddings described in Experiment 1. We concatenated the adjective and noun embeddings for each phrase (d=600). For the **BERT** embeddings, we used the CLS token, which acts as a combined representation of the adjective and noun, rather than concatenating the word-level embeddings. The CLS token is a classification token that is required at the beginning of every sentence input into BERT. It is understood as a sentence-level representation of the input, generated with classification tasks in mind (Devlin et al., 2018; Munikar et al., 2019). As it constitutes a broad representation of the meaning of multi-word inputs, we used the CLS token as the entire adjective-noun representation in our analyses. For our contextually prompted condition, we input phrases such as "the brightness of dark charcoal" into BERT and extracted the CLS token for the entire phrase. In contrast, for our context-free condition, we input only the adjective-noun pair, for example "dark charcoal" and extracted the CLS token. This meant that our BERT embeddings always had the same dimensionality (base: d=768; large: d=1024). It was not possible to average context-free embeddings over multiple sentence contexts (as in Experiment 1) because many of the adjective-noun phrases did not appear in the one Billion Words Benchmark corpus. Finally, we again used a one-hot encoding scheme to act as a baseline model. Inputs to this model consisted of one unit for each adjective-noun phrase.

Model

We again used 10-fold cross validation to evaluate our models. In Experiment 1, each test fold contained a subset of nouns that the model was not trained on so that we could test generalisation to these novel nouns. In the present investigation, we were instead interested in how accurately embeddings could predict the brightness of novel adjective-noun combinations, having been trained on the same adjectives and nouns in different combinations. Accordingly, for each train-test split, we made sure that at least one version of each noun (dark-noun, light-noun or heavy-noun) appeared in the training set. This ensured that we were not testing on a previously unseen noun, but rather on a previously unseen adjective-noun combination (see for examples). All other model specifications were the same as our noun investigations, with hyperparameter tuning for the number of hidden units and number of epochs performed specifically for the adjective dataset.

Evaluation

Our evaluation procedures are similar to Experiment 1. We again initialised 10 different models with randomised starting weights to avoid the effect of random noise. Each of the 10 models was trained and tested according to the 10-fold cross-validation scheme described for Experiment 1 (see Appendix for standard deviations of performance across the models). We obtained a single prediction for each adjective-noun combination by averaging predictions from the 10 models. We evaluated performance of the different embeddings, separated by adjective, using mean squared error (MSE) and R^2. As such, our metrics indexed the ability to predict brightness across the nouns when paired with the same adjective. Additionally, we ran statistical tests on our comparisons of interest, outlined below. We used the Wilcoxon signed-rank test, which is a paired-samples, non-parametric test, comparing squared errors for the "light" and "dark" adjective-noun phrases.

- Word2Vec vs context-free BERT
- "Brightness" contextually prompted BERT vs "colour" contextually prompted BERT
- Context-free BERT vs contextually prompted BERT

The best-performing embedding out of the prompt comparison (i.e., brightness prompt vs. colour prompt) was selected for the context-free and contextually prompted BERT comparison above. We also present a qualitative evaluation of how well embeddings capture the non-additive effect of adjective brightness on noun brightness.

4.4.2. Results

We evaluate the adjective investigations in a similar way to the noun investigations, however we report the model performance separated by adjective (see Figure 4.8 for an overview).

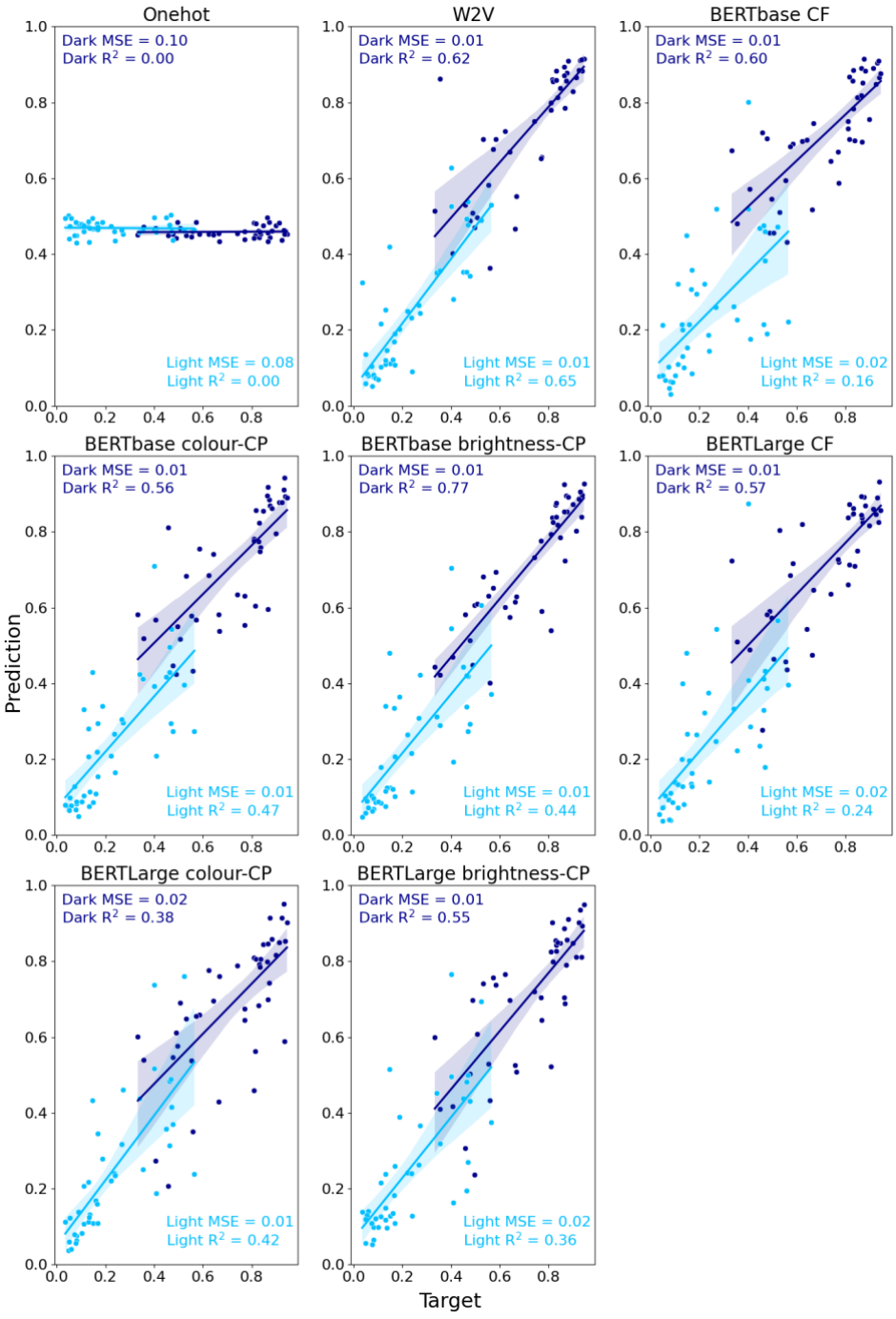

Figure 4.8. Predicted vs target brightness values for the S&T-S adjective-noun pairs (dark items = dark blue; light items = light blue). Shaded areas indicate 95% confidence interval for regression. CF= context-free, CP= contextually prompted.

In general, all of the embeddings could predict the brightness of novel adjective-noun phrases well. We found that the Word2Vec embeddings performed best overall (Dark: MSE= 0.01, R^2= 0.64; Light: MSE= 0.01, R^2= 0.70). Contrary to our predictions, in the analysis of our decontextualised embeddings, we found that Word2Vec embeddings were significantly better at perceptual prediction than the context-free BERT embeddings (base: p= 0.0001; large: p= 0.0007). For our comparison of prompt effectiveness, we found no significant differences between "the colour of" and "the brightness of" prompts (base: p= 0.17; large: p= 0.68). As such, for the contextual

analysis, we selected “the brightness of” prompt as it generally had higher R^2 values. We found that the contextually prompted BERTbase embedding performed significantly better than the context-free embedding (p= 0.02). However, there was no significant difference between the two for BERTLarge embeddings (p= 0.84).

One of the key aspects of the adjective-noun S&T-S dataset is the non-additive effect of adjectives on brightness ratings. This “flexible modulation” is shown in the top-left panel of Figure 4.9 which plots the human ratings for adjective-noun brightness (y-axis) as a function of noun brightness (x-axis). Adjectives strongly modulate the brightness of nouns that fall in the middle of the spectrum (e.g., “paint”), while they

have less effect on nouns with more extreme brightness values (e.g., “charcoal”). We can see a similar pattern (curvature of datapoints) for the word embeddings.

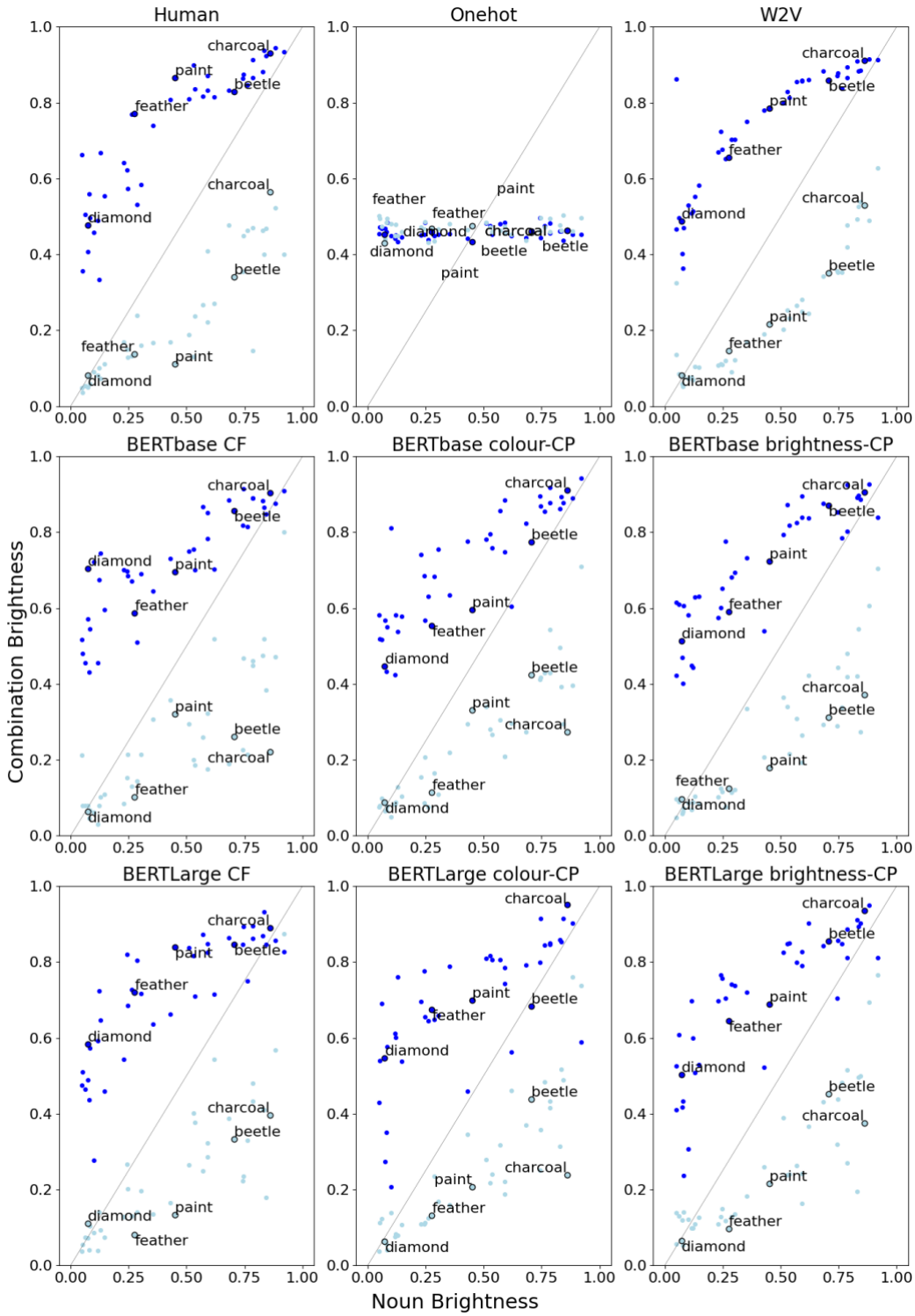

Figure 4.9. Noun vs combination brightness values for the S&T-S adjective-noun pairs (dark items = dark blue; light items = light blue) with human ratings (first subplot). CF= context-free, CP= contextually prompted.

4.4.3. Discussion

Overall, we observed good performance for perceptual prediction on unseen adjective-noun combinations for the pre-trained word embeddings, but not for the one-hot baseline. We also found that embeddings mimic the non-additive "flexible modulation" evident from human data, which suggests that LLMs can capture complex aspects of conceptual combination. However, here we found that Word2Vec embeddings outperformed context-free BERT embeddings, in contrast to our Experiment 1 findings. It is possible this is because the adjective and noun are represented separately in our Word2Vec inputs, as these were a concatenation of the two constituent embeddings, whereas our context-free BERT embeddings are a single, blended representation of both words. We found that the addition of context had a limited impact on perceptual performance, with modest evidence that including a prompt towards the feature could uncover additional feature-related information.

4.5. General Discussion

In this study, we investigated the ability of embeddings from Distributional Semantic Models and Large Language Models to represent perceptual information, comparing embeddings with different levels of contextual constraint. In Experiment 1, we compared contextualised and decontextualised word embeddings on their performance on perceptual prediction for the brightness and shape associated with nouns. In general, we found very good prediction of shape and more modest prediction of brightness, suggesting that language models do capture some perceptual aspects of meaning through exposure to linguistic information alone. Shape was generally better predicted than brightness, replicating previous findings (Chersoni et al., 2021; Turton et al., 2020, 2021). In our novel investigation of context, we found no advantage for contextualising the embedding with the desired perceptual feature (e.g., "the brightness of charcoal"). In fact, this most often led to worse performance. For Experiment 2, we explored whether word embeddings flexibly represent the modulation of perceptual properties that occurs when nouns are modified by adjectives (e.g., "dark charcoal"). This work extends previous findings to focus on how conceptual features from multiple inputs interact in multi-word expressions. Overall, the embeddings were successful in predicting the properties of novel adjective-noun combinations, and we found limited evidence for the effectiveness of contextual prompting on perceptual feature prediction. We also found that the non-additive effect of adjectives on noun brightness was represented within the embeddings, which mimics a key qualitative characteristic of the human dataset (Solomon & Thompson-Schill, 2020). These results have implications for understanding the degree to which aspects of embodied perceptual experience are coded in the statistics of language.

We found generally good prediction of perceptual features from the different sets of pre-trained word embeddings tested. This was even the case for brightness prediction, which is a feature that previous studies found to be poorly predicted (Chersoni et al., 2021; Turton et al., 2021; Utsumi, 2020). One caveat to note is the difference in performance when predicting brightness for the two datasets. Our results suggest that performance may depend critically on the set of concepts used to test

predictions. Previous studies relied heavily on the Binder dataset, which covers a wide range of concepts. We found better performance for brightness predictions when we used the S&T-S dataset, which contains a more tailored set of concepts for which brightness is a relevant feature. Solomon and Thompson-Schill (2020) curated their dataset to include concepts that represented the spectrum of brightness values. We found that this resulted in better prediction performance in contrast to the Binder dataset, which included many concepts where brightness was not a salient feature. As such, it may be important to use data that cover the spectrum of perceptual feature ratings when assessing the representational ability of word embeddings.

In addition, our findings suggest that shape is more strongly represented in word embeddings than the brightness of a concept, replicating previous results (Chersoni et al., 2021; Turton et al., 2020, 2021). This could be in part because the degree to which a concept has a shape is strongly related to the degree to which the concept is concrete or abstract. More abstract concepts (e.g., "government") reliably have lower ratings compared to more concrete concepts (e.g., "table"; see Figure 4.6). This relationship does not hold for brightness, where some abstract concepts (e.g., "summer") can be rated highly and many concrete concepts are neutral. As such, it is possible that shape prediction was aided by this feature's correlation with concreteness, which is highly salient both psychologically and linguistically (Barsalou et al., 2018; Paivio, 1990). However, better performance does not seem to stem simply from an effect of bootstrapping from concreteness information, as we observed better prediction performance for shape even when only concrete nouns were included.

In terms of the differences between word embeddings, we found that the context-free BERT embeddings, which were averaged across many contexts, generally outperformed Word2Vec embeddings in predicting perceptual ratings in Experiment 1. This replicates a previous finding and suggests that probing word embeddings across multiple contexts can lead to more robust representations than static embeddings (Bommasani et al., 2020; Turton et al., 2021; Vulić et al., 2020). From our novel investigations with the addition of a contextual prompt, we found that the contextually prompted BERT embeddings did not result in better performance than our context-free BERT embeddings. In some circumstances, we actually found worse performance. This does not align with what we observe in humans, where context has powerful effects in shaping how particular concepts are retrieved and interpreted (see Yee & Thompson-Schill, 2016 for a comprehensive review). In one apposite example, Bermeitinger, Wentura and Frings (2011) gave participants a task which drew attention to the shape of concepts, interspersed with a semantic priming task. This resulted in greater priming for concepts where shape was a relevant feature, compared to less shape-relevant concepts. This type of semantic facilitation based on the interaction between the context and a concept's features is a robust finding using both behavioural and neuroimaging methods (Hoenig et al., 2008; Hoffman et al., 2018; Kuhnke et al., 2020; Tabossi & Johnson-Laird, 1980; Van Dam et al., 2010; van Dam et al., 2012; Yee et al., 2012). Findings like these indicate that the perceptual features humans activate upon processing a concept are strongly influenced by the recent context. However, contextual prompting does not appear to shape the expression of feature-specific information in Transformer-based LLMs in the same way. Our results suggest that

specific perceptual prompts may not engineer representations that better reflect that particular property and that context-free embeddings (aggregated over many contexts) are more effective. This may be one way in which the semantic representations extracted from language models differ from our understanding of semantic representations in the human brain. Future work that considers a wider range of features and prompts would give greater insights into the ways in which context influences the representation of different semantic features in word embeddings. It could also be possible that newer, larger LLMs are better able to capture context than BERT, which is not as well-suited to this type of prompting. As such, future explorations of how different model architectures handle this task may also be of interest. Between the two versions of the BERT model, it is interesting to note that BERTLarge did not consistently outperform BERTbase embeddings. BERTLarge is a substantially larger model, with more than three times as many parameters as BERTbase, and was reported to outperform BERTbase on a range of language processing tasks (Devlin et al., 2018). However, others have suggested that the BERT models were significantly under-trained (Liu et al., 2019) and it is possible that with greater training, BERTLarge would have consistently overtaken the smaller model in our perceptual comparisons.

In Experiment 2, we explored the ability to predict perceptual feature values from word embeddings for novel adjective–noun pairs, having had experience of each of their constituents. We found that the Word2Vec embeddings generally outperformed the context-free BERT embeddings. Here, our Word2Vec embeddings were a concatenation of the adjective and noun embeddings, while for our context-free BERT embeddings we took the CLS token to represent the entire adjective-noun phrase. It is possible that having distinct representations of the adjective and noun resulted in the better performance for Word2Vec embeddings in this task. Nevertheless, the BERT embeddings did perform well on this task, which suggests that an integrated phrase-level representation (i.e., CLS token) also carries perceptual information. We also found a limited impact of contextual prompting on predictions, as our contextually prompted BERTbase embeddings performed significantly better than our context-free BERTbase embeddings. However, this relationship did not hold for the BERTLarge embeddings. We take this as modest evidence that the provision of a contextual prompt can uncover further feature-related information. We also found evidence of the non-additive nature of adjective-noun brightness reflected in the embeddings. This characteristic refers to the variable way in which modification with an adjective impacts the brightness ratings of nouns (Solomon & Thompson-Schill, 2020). For example, a mid-brightness concept, such as “paint”, has greater variation in brightness ratings across adjective combinations, than an extreme-brightness concept, such as “charcoal”. This finding suggests that the type of flexible modulation found in how humans represent concepts is also reflected in word embeddings. Of note, we want to highlight the ambiguity here between the senses of “light” as both the antonym of “dark” and the antonym of “heavy”. It is possible that this impacted the comparisons we make between the unmodified noun predictions and the modified phrases. Indeed, polysemy remains a challenge for compositional distributional approaches to adjective modification precisely because of idiosyncrasies in the compositional process (Baroni, 2013; Baroni & Zamparelli, 2010; Boleda, 2020). For example, the different senses of “light” could interact with the same noun, such as “coffee”, to differing degrees. Future work could

investigate the polysemous nature of perception-relevant adjectival modification by comparing contextualised word embeddings that reflect different senses such as in "light".

Overall, our study converges with previous investigations in suggesting that a surprising degree of perceptual information can be transferred through linguistic content alone. Utsumi (2020) suggests that language is a realisation of experiences in the real world, arguing that the type of statistics used by language models implicitly involve conceptual knowledge from direct experience. Our findings support this view, which is most compatible with a unified account of cognition, where processing integrates embodied and symbolic elements (Andrews et al., 2014; Louwerse, 2011, 2018). Accounts such as the Symbol Interdependency Hypothesis emphasise the use of both symbol-to-symbol mappings, as well as symbol-to-world mappings (Louwerse, 2011). The current work adds to the body of literature indicating that semantic representations learnt using distributional methods frequently align with those based on reports of perceptual experience (Chersoni et al., 2021; Lucy & Gauthier, 2017; Sommerauer & Fokkens, 2018; Utsumi, 2020). Additional evidence for a dual representational system is reviewed by Bi (2021), who highlights behavioural and neuroimaging studies of colour knowledge on both those with visual experience, and those without (i.e., congenitally blind participants). Evidence from neural decoding demonstrated a non-sensory, language-derived system of colour knowledge in both sets of participants, with an additional sensory-derived representation present for those with visual experience. This suggests that humans can and do acquire perceptual knowledge through language to some degree. This evidence is more in line with a weak embodiment approach, such as that put forward by Dove (2014), who argues that language itself is embodied and interacts with other embodied systems (i.e., perception and action). Another relevant account is the Linguistic and Situated Simulation (LASS) theory (Barsalou et al., 2008; Santos et al., 2011), which proposes that lexical–semantic processing involves the early activation of linguistic information and is supported by later, embodied processes that simulate relevant sensorimotor experiences.

Possible directions for future research include investigating how different types of semantic features are encoded in language for both concrete and abstract concepts. Specifically focusing on meaning within LLMs, Piantadosi and Hill (2022) argued that meaning is not just borne out of grounding, but rather in how concepts relate to each other. They claim that this mapping of interactions between concepts is common to both humans and machines. Exploring the mechanisms behind how LLMs gain this kind of representational ability would be an interesting direction for future work. Another possible direction is to investigate the extent to which perceptual properties of interactions are represented in models. While we have explored the nature of adjective and noun interactions, there are many interactions involved in a myriad of conceptual combinations, such as noun-noun compounds, which warrant further investigation (Coutanche et al., 2019). Moreover, work that considers different implementations in the extraction and testing of word embeddings would address one of the limitations of the current work. For example, this could include the use of more elaborate prompts when contextualising towards a specific feature, such as prompts that convey events in

the form of transitive-verb clauses. Further research in this area would allow for greater analysis of the generalisability of the current findings.

In conclusion, our current work adds to recent literature that probes the relationship between word embeddings and human semantic ratings. We replicated previous results that demonstrated generally good performance for the prediction of some perceptual features, namely shape, while brightness was less well represented. In our novel contributions, we found that the addition of a contextual prompt had limited improvements on the representational ability of word embeddings for perceptual prediction. Moreover, word embeddings do reflect some of the flexible modulation of perceptual features that occurs in semantic compositions, in particular, modification with an adjective. Future research could focus on generating more specific context prompts and extending this to other features. This may reveal further insights into how linguistic context impacts engineered semantic representations.

Chapter 5

Flexibility in conceptual combinations: a neural network model of gradable adjective modification

5.1. Abstract

Our ability to combine simple constituents into more complex conceptual combinations is a fundamental aspect of cognition. Gradable adjectives (e.g., “tall” and “light”) are a critical example of this process, as their meanings vary depending on the noun with which they are combined. For example, a *dark diamond* is less dark than *dark charcoal*. Here, we investigate how a neural network encodes the flexible nature of gradable adjectives in adjective-noun pairs, using the perceptual feature of brightness as a test case. We trained a neural network to predict human brightness ratings for unmodified nouns and adjective-noun pairs and assessed its ability to generalize to untrained combinations (e.g., “light paint” vs. “dark paint”). We also explored how this information is encoded. We found that flexible learning of gradable adjectives was possible, with neural networks first making predictions based on the adjective alone, and then modulating these with information from the noun later in learning. We also found that model outputs mimicked the kind of non-additive feature modulation present in human data. Our results have implications for understanding how semantic composition occurs and generate testable predictions for future work.

5.2. Introduction

Conceptual combination refers to the ability to construct complex concepts from simpler constituents. For example, even if you have never encountered combinations such as “sand gun” and “robin eagle”, you are able to infer what such concepts may be by relying on the semantics of the constituent words (Costello & Keane, 2000; Coutanche et al., 2019). This process is highly dependent on context, with varying outcomes on the semantic representation of the combined phrase. Understanding how conceptual combinations are constructed may be able to help our understanding of conceptual representations in general (Coutanche et al., 2019). Previous theories of conceptual combinations from cognitive science have posited two mechanisms: attributive and relational. An attributive process is where an attribute of a word is assigned onto another, such as “zebra clam” to describe a clam with stripes, whereas a relational process concerns the inference of the relationship between two words, such as “floor television” to describe a television standing on the floor (Estes, 2003; Wisniewski, 1997; Wisniewski & Love, 1998).

What makes adjective-noun combinations such an interesting use-case is their reliance on context in the relation between adjective and noun (Asher, 2011; Boleda et

al., 2013), which is the focus of the current study. Adjectives, which are descriptive words that modify the word they attach to (Dixon & Aikhenvald, 2004), are able to modify meaning in multiple ways. Adjective-noun pairs that contain a relative gradable adjective (e.g., "tall penguin") are particularly context-dependent. Solt (2019) highlights that these adjectives are only understood in relation to a comparison class. For example, the meaning of adjectives such as "tall" and "light" are understood against the contextual standard for the group of objects that they modify. To explain further, you can use "tall" to describe someone of above average height, and also for a building such as The Shard. Here, the actual height denoted by "tall" is different between the examples as the same adjective can have different effects depending on the noun with which it is paired. As such, the meanings of gradable adjectives are inherently dependent on their context. In general, the adjectival modification of nouns presents an interesting challenge for distributional language models due to the highly variable nature of semantic composition (Asher, 2011; Boleda et al., 2013). We argue that adjective-noun pairs which include a gradable adjective present an even greater challenge, and that insights from neural network modelling could help us better understand this composition process.

The current study attempts to computationally model this flexibility in conceptual combinations, focusing on adjective-noun pairs. We trained a feedforward neural network to predict brightness ratings for adjective-noun pairs (e.g., "light paint") and tested its ability to generalize to unseen combinations. We presented information about both the unmodified concepts (here represented by a neutral adjective-noun pair condition) and their dark and light combinations. The brightness ratings were taken from Solomon and Thompson-Schill (2020), where humans were asked to rate the darkness of concepts for both unmodified nouns and adjective-noun pairs. We compared our model's performance against the generative models presented by Solomon and Thompson-Schill (2020) and present qualitative explorations into how our model performs this task. We found that our model can learn to predict brightness values for adjective-noun pairs and can successfully generalize to unseen adjective-noun combinations, performing at a similar level to Solomon & Thompson-Schill's Bayesian model, while outperforming simpler additive and multiplicative models. Moreover, we found that our model first learns information about adjective brightness, then begins to combine this additively with knowledge of noun brightness, and only later learns to combine noun and adjective knowledge in the non-additive fashion observed in the human data. Concepts that are more ambiguous with regards to their brightness (e.g., "paint") were also learnt later in training, compared to those that were not (e.g., "charcoal").

To emphasize, the current work is focused on the question of *how* models encode this information, rather than concerns about performance, as this can provide novel insights and generate further hypotheses about the process of semantic composition. It has recently been suggested that neural networks are a promising method for capturing both the systematic and idiosyncratic aspects of language, due to the lack of constraints imposed on the internal representations used when mapping inputs to outputs (Rabovsky & McClelland, 2020). Thus, it is possible that these types of models would be useful in modelling the flexibility of gradable adjectives, which

combine systematic constraints with idiosyncratic item-related biases to construct a meaningful interpretation.

5.3. Related Work

Much of the computational work on semantic composition has implemented vector- and matrix-based compositional functions to represent combined concepts (Baroni & Zamparelli, 2010; Hartung et al., 2017; Mitchell & Lapata, 2010). Hartung et al. (2017) aimed to model adjectival attribute meanings using word embeddings. Attribute selection is the task of predicting the hidden attribute meaning that is expressed by an adjective-noun combination; for example, the difference between understanding that "hot summer" relates to the temperature of the combined concept, whilst a "hot debate" relates to the passion surrounding the topic. By making use of a dataset with attribute annotations, they found that weighted combinations of adjective and noun embeddings could accurately predict the attribute described by a phrase, outperforming predictions from either the adjective or noun alone (Hartung et al., 2017). While their findings on how meaning is represented in adjective-noun pairs is of interest, the investigations in Hartung et al. (2017) only predict which attribute can be assigned to the adjectival modifier (e.g., weight, brightness, speed), rather than the magnitude of the modifier's influence. Thus, the question of how to build flexibility into computational representations of gradable adjective-noun pairings remains.

Shwartz and Dagan (2019) identified six tasks associated with compositional phenomena and tested how well a range of word embeddings could accurately reflect the lexical composition process. Overall, they found that contextualized word embeddings performed better at the tasks, compared to static embeddings. However, while they exhibit similar performance to humans at recognizing meaning shifts, performance was much lower for tasks that required a representation of implicit meaning. This highlights the difficulty distributional models have in representing the meanings of phrases, especially those with contextually-dependent interpretations (Asher, 2011; Boleda et al., 2013; Shwartz & Dagan, 2019).

Solomon and Thompson-Schill (2020) have recently attempted to model flexibility in conceptual combinations. They used a three-pronged approach, incorporating behavioural, computational and neuroimaging methods to explore conceptual structure and the neural regions that support the flexible use of features. The authors focused on the level of perceptual brightness conveyed by adjective-noun pairs. They introduced a construct, feature uncertainty, which reflects the entropy associated with a concept's brightness (Shannon, 1948). In their behavioural experiments, human participants rated the brightness of 45 modified and unmodified concepts on a scale from 0 (light) to 50 (dark) (see "Human" plot in Figure 5.1.). They found that brightness ratings were influenced by both the adjective and noun. For example, the ratings for "light feather" were lighter than those for either "dark feather" or "light charcoal". It was also apparent that the degree to which the adjective modulated brightness was not constant across nouns. For example, some of the concepts had large differences between their light and dark modified forms (e.g., "paint"), whereas for

other concepts, this difference was much smaller (e.g., “white”). The authors found that the flexible modulation of brightness across concepts correlated with their construct of feature uncertainty: the degree of adjectival modulation was greatest for objects of moderate brightness (e.g., “paint”, “slippers”; which were assumed to have the greatest feature uncertainty), and smallest for objects with more extreme values of brightness (e.g., “snow”, “charcoal”). As such, their data suggests a predictable, but non-additive relationship between the expected brightness of an adjective-noun pair and the brightness of its adjective and noun constituents.

The authors also implemented a number of generative models for brightness prediction. They incorporated two baselines, where the predicted brightness of the combination was just the brightness of either the noun or adjective, respectively. They also included an additive model, which predicted combination brightness through a weighted sum of adjective and noun brightness; a multiplicative model, which predicted combination brightness through a scaled product of adjective brightness and noun brightness; and a Bayesian model, which generated predictions through a product of Gaussian brightness distributions for the adjective and noun, fit on the response frequencies from the behavioural judgement task. They found that the Bayesian model significantly outperformed the other models. As the Bayesian model was the only model to incorporate information on feature uncertainty (i.e., the variability in brightness ratings for each object), the authors argued that feature uncertainty was critical for capturing the patterns of feature modulation in the human judgements (Solomon & Thompson-Schill, 2020).

In the present study, we investigated how a simple neural network learns to predict the brightness of adjective-noun concepts. The network was trained on a subset of Solomon and Thompson-Schill’s (2020) adjective-noun brightness ratings and tested on its ability to predict brightness for unseen, novel combinations of adjectives and nouns. This work represents an advance on previous work in two ways. First, existing models provide accounts of how adjective and noun information combines in a mature semantic system but are largely silent on how this ability is acquired. As neural networks learn to perform tasks incrementally through training, they provide an opportunity to investigate how representations emerge and what developmental stages are involved (Frank et al., 2019; Rogers & McClelland, 2004). Second, unlike the Bayesian model proposed by Solomon and Thompson-Schill, our simulations included no notion of feature uncertainty. This allowed us to test whether the construct of feature uncertainty is necessary to account for non-linear effects of adjectival modification.

5.4. Methods

All associated code and data can be accessed here: https://osf.io/ptqnu/

5.4.1. Dataset

The dataset from Solomon and Thompson-Schill (2020) consists of averaged human ratings from a behavioural experiment, where human raters were asked to rate the

brightness of unmodified nouns (e.g., “coffee”) and modified adjective-noun pairs (e.g., “light coffee” vs “dark coffee”) for 45 concepts. The original dataset from Solomon and Thompson-Schill (2020) can be accessed here: https://osf.io/7uwn9/. Two separate groups of participants (*n*=100; *n*=199) rated the brightness of the unmodified nouns and the brightness of the adjective-noun combinations. The brightness ratings were on a scale from 0 to 50, with 0 representing light and 50 representing dark.

We used the averaged ratings of the brightness of the unmodified concepts and the averaged ratings of the brightness of the combined concepts (for example, the concept “black” had an unmodified rating of 47.83, while “dark black” had a rating of 49.61 and “light black” a rating of 37). We transformed these ratings to a scale from 0 to 1, with 1 now representing the dark end of the spectrum. To standardize our inputs to our model, we appended a brightness-agnostic adjective (“neutral”) to the unmodified concepts. Therefore, we had three versions of each concept: dark, light and neutral, resulting in 135 items in total. To generate our model inputs, we created a one-hot encoding of both the noun and the adjective, and then combined these to form a representation of the adjective-noun pairs.

5.4.2. Model

We implemented a feedforward neural network architecture in PyTorch (Paszke et al., 2019). The network consisted of three layers, with one hidden layer. The input layer consisted of 48 units, representing the 45 nouns and 3 adjectives. The hidden layer had 30 units, while the output layer consisted of 1 unit, which represented the model’s brightness prediction. Between the linear layers, we included a Rectified Linear Unit (ReLU) activation function (Agarap, 2019), while we used a Sigmoid activation function between the hidden and output layers in order to transform the model prediction between 0 and 1, and thus be comparable to our scaled brightness ratings.

We set a range of hyperparameters, with some values optimized through grid search (see Training), and others taken from a study with a similar goal of representing flexibility in semantic concepts (Hoffman et al., 2018). As such, our model had a *bias*=-2, *momentum*=0.9, and *weight decay*=10^{-6}.

5.4.3. Training

Due to the limited size of our dataset, we implemented k-fold cross validation (*k*=10) in order to maximize the utility of our data (Arlot & Celisse, 2010; Fushiki, 2011; Geisser, 1975). We chose *k*=10 as it has been widely used across the machine learning literature (Arlot & Celisse, 2010; Marcot & Hanea, 2021; Nti et al., 2021). We split our dataset into train and test sets, with approximately 122 items in train and 13 items in test. We ensured that the nouns present in the adjective-noun pairs in the test set were also present in a different combination in the train set. For example, if “dark charcoal” was a test item for one of our folds, then we confirmed that the train set contained at least one “charcoal” item, such as “neutral charcoal”. We fed the input items to the model in batches, with a batch size of 14. We performed hyperparameter optimization using

nested k-fold cross validation (*n*=3), such that our training set was further split into three sets, with one of these sets used as a validation set. We implemented grid search, whereby we optimized on learning rate, the number of hidden units and the number of epochs to train for (Liashchynskyi & Liashchynskyi, 2019). We evaluated our grid search using the negative mean squared error (MSE), which resulted in optimal parameters of *learning rate*=0.3, *number of hidden units*=30 and *train time in epochs*=125. In our final models, training ran for 125 epochs, with our model weights optimized through stochastic gradient descent (Amari, 1993).

5.4.4. Evaluation

To evaluate our model's predictions on the unseen adjective-noun pairs, we used mean-squared error, comparing the model brightness predictions for the unseen adjective-noun pairs against the ground-truth, i.e., averaged brightness ratings from human participants, and R^2. As such, during training, the model acquires knowledge about the typical brightness of a range of objects and is shown how the two adjectives ("dark/light") modulate brightness for some, but not all, of these objects. It is then tested on the combinations that were not provided during training. Thus, we tested the model's ability to acquire knowledge about how dark/light adjectives modulate the expected brightness of objects, situated along the brightness spectrum, and then to generalize this knowledge to novel adjective-noun combinations.

5.5. Results

We trained 10 models initialized with different random weights. Each model was trained for 10 iterations using k-fold cross-validation. All results below are averaged over the 10 models and only include performance on unseen adjective-noun combinations.

5.5.1. Model Performance

In Figure 5.1, we plot the model predictions for the held-out combined concepts alongside the human ratings, after training for the full number of epochs. These are separated by adjective, with annotated examples taken from Solomon and Thompson-Schill (2020). The brightness of the unmodified concept is plotted on the x-axis, against the predictions of combination brightness on the y-axis. For instance, if we take 'paint' as our example, Figure 5.1 shows that the model predictions for the modified noun have the same x-axis value (i.e., compare the x-axis values for model-derived predictions for 'light paint' (light red triangle) and 'dark paint' (dark red triangle)), while the y-axis values (i.e., combination brightness value) are different. Here, a value of 0 refers to the lightest possible object, while 1 refers to the darkest items. The grey line across the plots indicates the alignment of the combination brightness with the brightness of the unmodified concept. The model performed comparatively well in predicting the combination brightness of concepts after the full training procedure. Further, the model captured the three main features of the human data: (1) that the brightness of the adjective-noun pair is influenced both by the adjective and the noun, (2) that the degree

to which the adjective modulates the brightness varies across nouns and (3) that the largest modulations occur for nouns of moderate brightness.

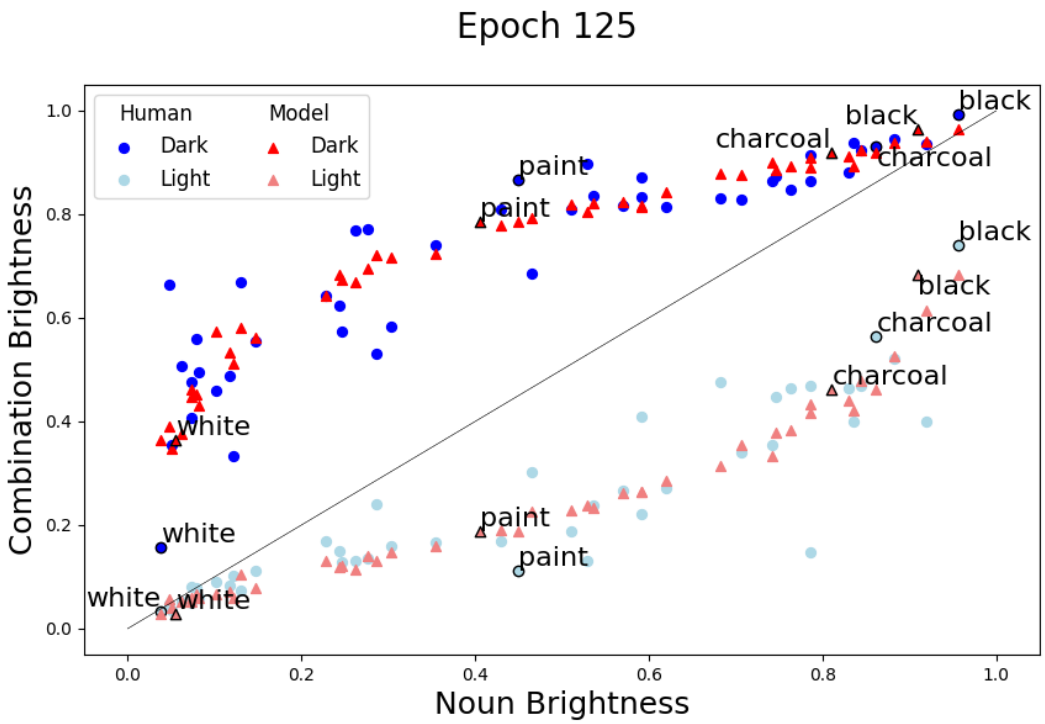

Figure 5.1. Human ratings and model predictions of combined brightness as a function of brightness for the unmodified concepts.

To investigate how the model evolved during training, we plot model predictions across a subset of epochs during the training procedure (see Figure 5.2). The model passes through a series of developmental stages. The model begins to cluster the combined concepts early on during training by making brightness predictions based on the adjective alone. The subplot of Epoch 4 highlights this clearly, with the "dark" items depicted in dark blue, and the "light" items depicted in light blue. The model is correctly predicting the difference between "dark" and "light" items but is insensitive to the noun brightness. However, by Epoch 10, the model acquires knowledge of noun brightness in order to assign a more accurate prediction (i.e., combination brightness begins to be influenced by noun brightness). Here, the model's predictions resemble the additive model from Solomon and Thompson-Schill (2020), in that the model is sensitive to the brightness of both the noun and the adjective, but the adjective modulates each noun's brightness to the same extent. This modulation becomes more flexible and noun-dependent in the later epochs, as demonstrated by the eventual non-linear curves in the Epoch 100 plot. Here, it appears that the model is gradually refining its predictions as it learns that the adjectives can have variable influences on different nouns, for example, a greater influence for concepts that fall in the centre of the brightness spectrum. Solomon and Thompson-Schill (2020) suggest that this feature of the human data is due to a greater amount of uncertainty for moderate-brightness nouns. However, there was no uncertainty in the inputs to our model—each adjective-noun combination was associated with a single, fixed brightness value. This suggests that the non-additive

modulation patterns in the human data can be explained without appealing to feature uncertainty.

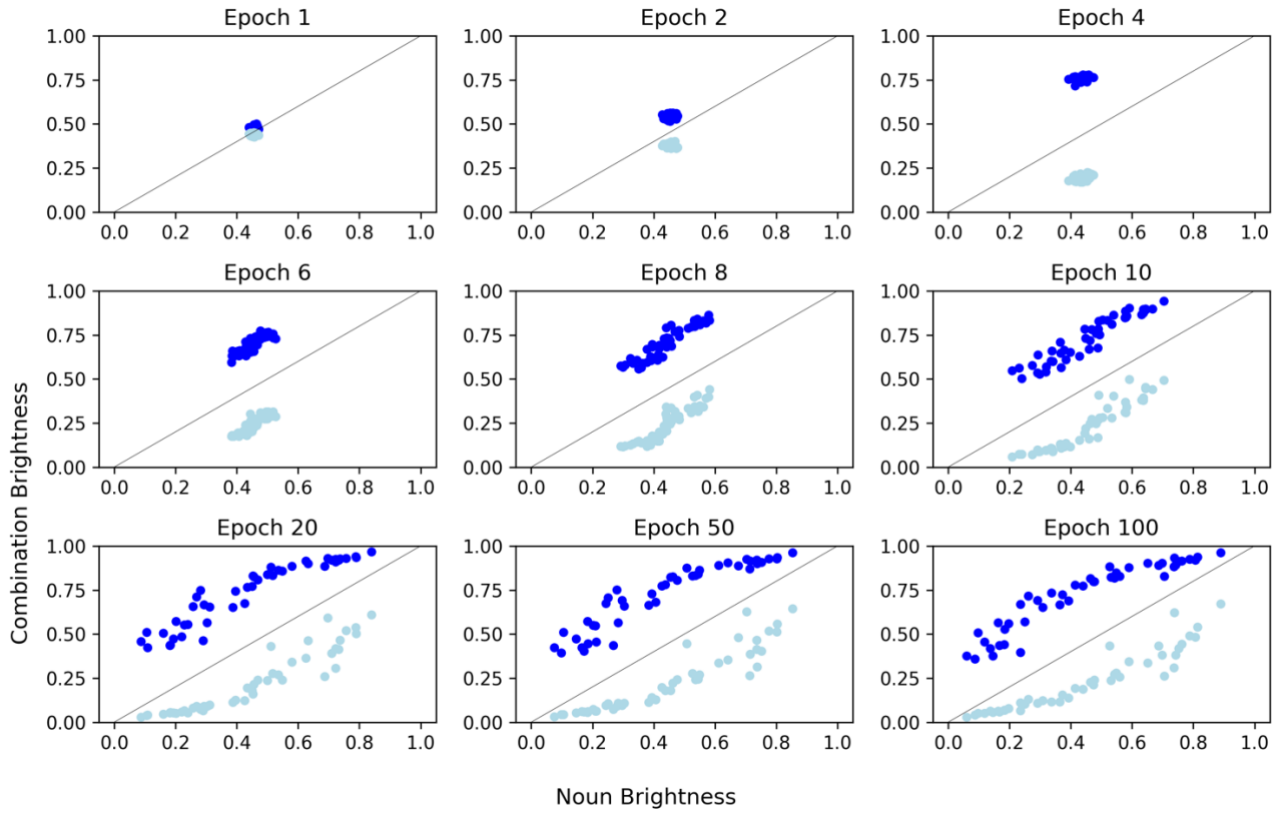

Figure 5.2. Model predictions of combination brightness during a subset of epochs.

5.5.2. Model Comparison

In order to ascertain how our neural network performed in comparison to the models presented in Solomon and Thompson-Schill (2020), we replicated their models using the data provided by the original authors. We also transformed the neural network predictions back to the original brightness scale (0=light, 50=dark) for comparability. We then evaluated the models' performances using mean squared error (MSE), R^2, and the standard deviation, and compared these across all models. Results can be found in Table 5.1, with results from our model depicted in bold. We also performed statistical analyses on the squared errors of the combinatorial models. A one-way ANOVA (analysis of variance) demonstrated that overall MSEs differed across the four models ($F_{(3, 176)} = 19.22$, $p < 0.01$). Pairwise comparisons revealed that the Bayesian and neural network models did not differ significantly in performance ($t_{(44)} = 0.08$, $p=0.94$). The neural network did significantly outperform both the additive and multiplicative models, however ($t_{(44)} = 2.24$, $p=0.03$; $t_{(44)} = 5.13$, $p < 0.01$).

Table 5.1. Model comparisons. Bold indicates our implementation; all other implementations are from Solomon & Thompson-Schill (2020). % change indicates the

percentage difference in MSE between our neural network and all other implementations.

Model	MSE	SD	R^2	% change
Adjective	258.63	25.0	0.00	93.7
Noun	207.30	14.88	0.09	92.1
Additive	29.57	17.76	0.87	44.5
Multiplicative	80.24	20.47	0.65	79.6
Bayesian	16.87	14.85	0.93	2.7
Neural network	**16.41**	**15.13**	**0.93**	

5.5.3. Learning Trajectories

To understand the mechanisms that supported learning, we ran qualitative explorations into the model's performance. We first outline our investigations into the learning trajectories of the annotated examples across epochs. After, we discuss the activations of the hidden representations and present a cluster-based analysis using t-SNE (t-distributed stochastic neighbour embedding) (Maaten & Hinton, 2008).

To understand how our model performs on selected cases, we used the annotated examples shown in Figure 5.1. These contain concepts across the range of brightness ratings. Figure 5.3 depicts the model predictions of the combined concepts, separated by adjective. We plot these predictions across epochs on a log-scale to better demonstrate the distinction in predictions between the earlier and later epochs. Figure 5.3 shows that the predictions for the combined concepts become distinguishable by noun only later during training. This again highlights the clustering of combined brightness by adjective that dominates the model's initial predictions.

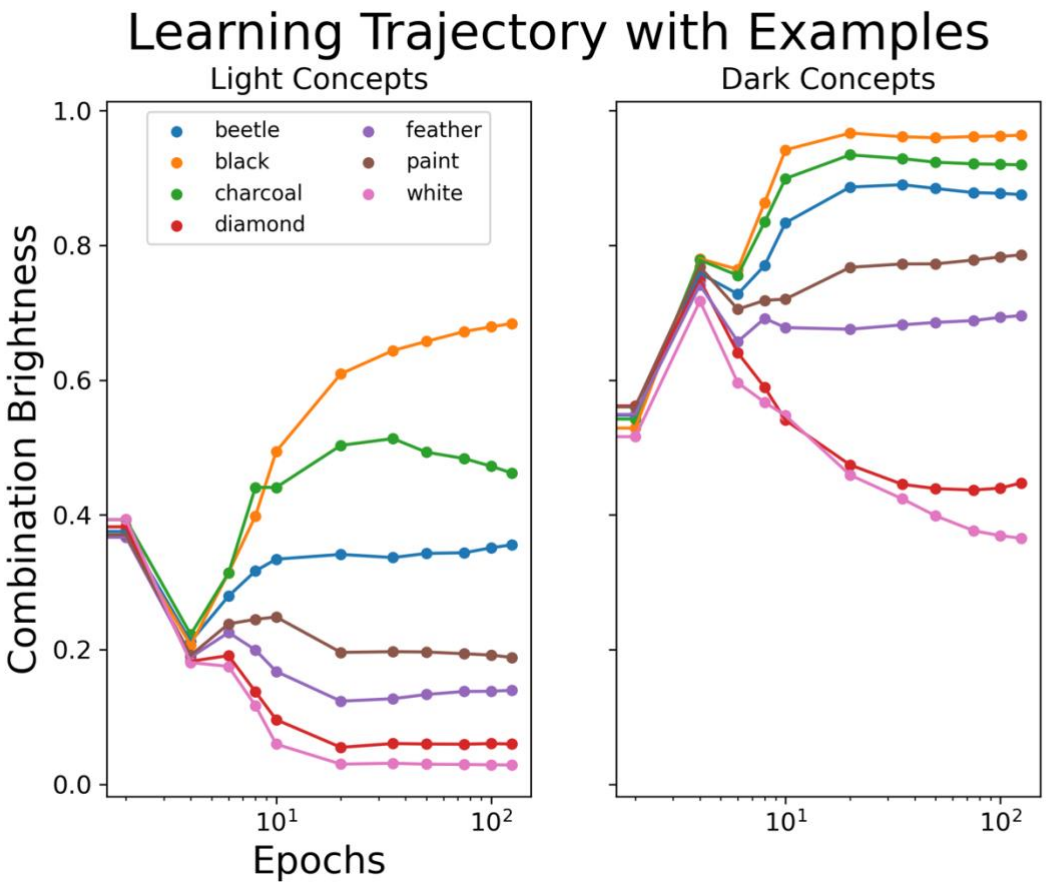

Figure 5.3. Model predictions for annotated examples over training. Model predictions of combination brightness across epochs (on a log-scale) for selected concepts, separated by adjective (light items on the left plot; dark items on the right plot).

We also investigated the error in model predictions for these annotated examples across all epochs (see Figure 5.4). Here, we define error as the numerical difference between the model predictions and the true combined brightness value. As such, negative values indicate that the model's predictions were darker than the true combined brightness value (i.e., closer to 1), whereas positive values represent model predictions that were lighter than the true combined brightness value (i.e., closer to 0). The large peaks in the error values are another indication of the model's predictions first assigning similar values to combinations with the same adjective. For example, the high error peak for "light black" (see orange peak in the left plot), compared with the high error peak for "dark white" (see pink peak in the right plot). This shows that the model is slowest to learn appropriate brightness predictions for concepts where the adjective and noun have contradictory brightness associations.

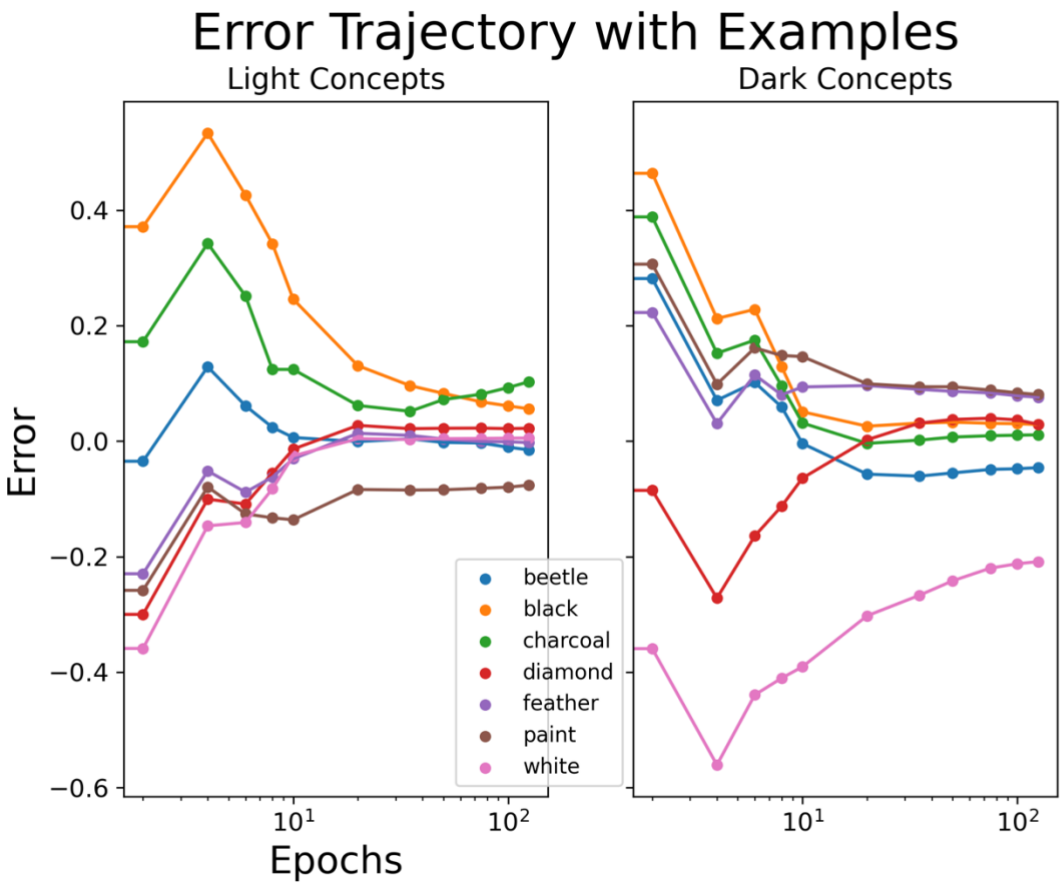

Figure 5.4. Error between model predictions and ground-truth during training. Prediction error of combined brightness predictions against true values across epochs (log-scaled) for selected concepts, separated by adjective (light items on the left plot; dark items on the right plot).

5.5.4. Hidden Representation Analysis

Finally, we analysed the hidden representations acquired by our neural network. We extracted the hidden activations for each item after training (epoch 125). We then performed dimensionality reduction using t-SNE to reduce the hidden activations from 30 dimensions to 2 dimensions. We ran this procedure for each of our folds to ensure that our output was consistent. Here, we only depict a representative figure from one of our 10 folds. In Figure 5.5, we can observe that the hidden activations of our items form two clusters based on the adjective, with light items represented by the circles, and dark items represented by the crosses. The darkness of the points refers to the unmodified noun brightness. In none of the investigations of the hidden activations did we find clusters based on the noun. This reinforces our previous findings that adjective identities are the dominant organizing principle for the model's representations.

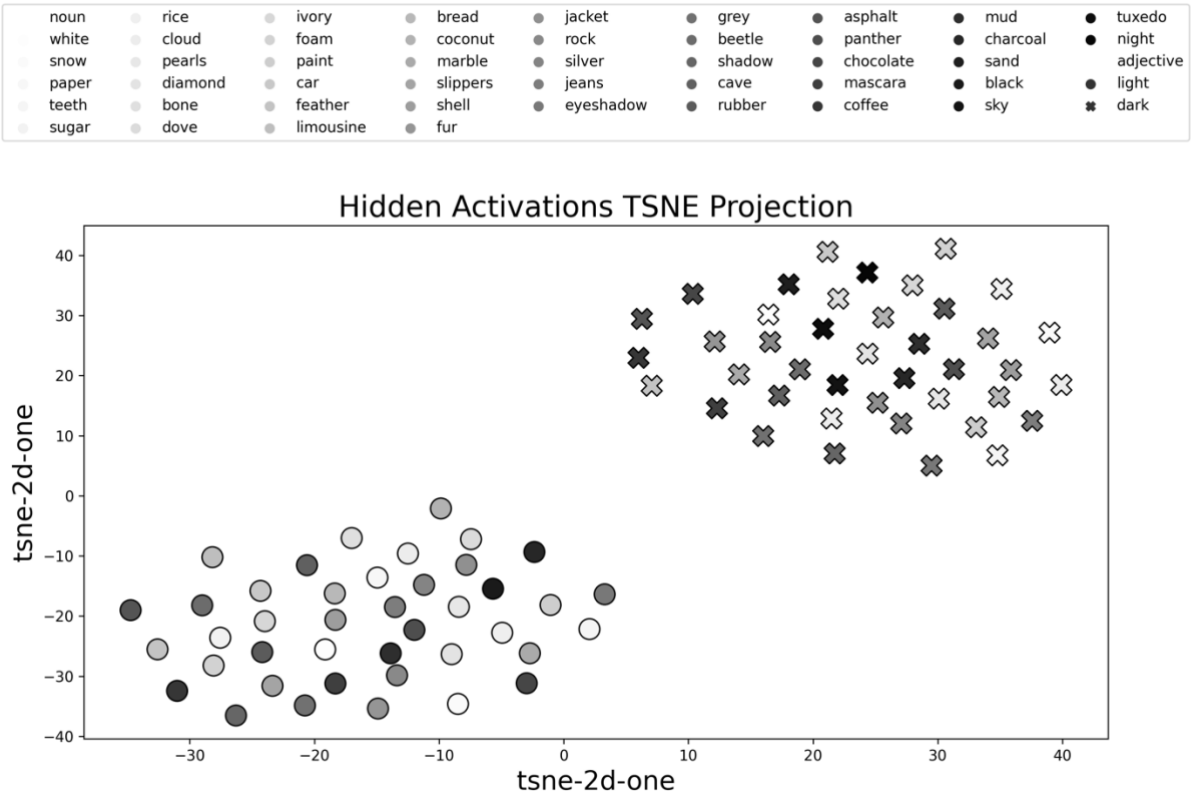

Figure 5.5. Hidden activations after 2D-TSNE reduction. Adjectives depicted by marker shape (circles represent light items; crosses represent dark items); nouns depicted through greyscale colour (ordered by unmodified brightness).

5.6. Discussion

In this study, we investigated how a neural network learns to encode the flexible nature of semantic composition with relative gradable adjectives. We trained a small neural network to predict brightness ratings of a range of concepts using both unmodified nouns and adjective-noun pairs. When tested with novel adjective-noun combinations, our model implementation performed as well as the Bayesian model presented in Solomon and Thompson-Schill (2020). While both models exhibit similar MSE and R^2, the Bayesian implementation is fit with a richer dataset (i.e., the distribution of ratings across individual participants), whereas our neural network achieves similar performance using only the mean ratings. As such, we argue that our neural network demonstrates an improvement over the Bayesian model due to the requirement for less training data. In addition, our neural network does not make use of the novel construct of feature uncertainty, which suggests that data on the uncertainty of a property is not needed to predict the influence of gradable adjectives on adjective-noun pairs.

Furthermore, the nature of neural networks allowed us to investigate exactly how the model predictions developed across training. Our investigations into the learning trajectories of specific examples revealed that the network first clustered the items by adjective. Later in training, the influence of the noun's brightness plays a role, with model predictions assuming an additive nature whereby the adjective modulates each

noun to the same extent. Towards the end of training, model predictions finally converge on a non-additive mapping, whereby the adjectives exert differential modulation of the combination brightness, depending on the nouns they are paired with. One question that emerges is whether children acquire knowledge about gradable adjectives in the same way. Previous research into the acquisition of gradable adjectives has demonstrated that children as young as 4 years old are able to interpret adjectives in a way that is sensitive to statistics of the object class they are applied to (Barner & Snedeker, 2008). However, there is little evidence on earlier stages of acquisition, so it is currently unknown whether children first develop adjective-based representations before this noun-specific information is incorporated. It was also found that children demonstrated an asymmetry in their mastery of compositional semantics with regards to positive and negative terms (e.g., better mastery of "tall", compared with "short") (see Smith et al., 1986). We did not find this asymmetry within our neural network, as predictions for both light and dark items appeared to develop similarly across epochs (i.e., first assign a blanket adjective prediction, then nuance by noun). It is possible this is because we did not provide any information as to the valence of the items. In other words, the model is not aware of which adjective corresponds to a positive or negative term in the real world. It is also possible that the age of acquisition (AoA) of positive and negative terms influences this asymmetry. One suggestion for further research would be to replicate this asymmetry in the acquisition of compositional semantics to better understand the mechanisms surrounding the influence of context on combined concepts. For example, a possible AoA influence could be introduced through focusing training on lighter items during earlier epochs.

A key question that arises from the findings of the current study is whether this approach extends to other conceptual properties. The use of brightness as our property of interest has some caveats, in the sense that it is strongly perceptually grounded. It has been demonstrated that perceptually salient cues are particularly important in the grounding of cognition (Barsalou, 2010). It is possible that less perceptually grounded and concrete properties, such as "expensive", are less amenable to this approach, especially considering the greater individual variability in rating more abstract concepts (Wang & Bi, 2021). The extension of this approach to other properties is, therefore, an interesting direction for future research. For example, further investigations with both different conceptual features of interest and adjectival types could assess whether the organizing patterns we observe here are general features of adjective-noun combination.

One limitation of the current approach is the simultaneous presentation of the adjective and noun representations to the model. As such, we were not able to investigate the impact of sequential presentation on compositional semantics. The use of sequential models would allow us to further investigate the mechanisms that support compositional semantics in spoken language. With more complex sequential models, such as a model trained to predict the noun following a presented adjective as well as its expected brightness, future research could also focus on the interplay between language and embodied perceptual predictions. This would enable further exploration into the nature of statistical influence on the acquisition of compositional semantics,

and thus, how the preceding context supplies comprehenders (whether human or artificial) with prior expectations that shape the semantic interpretation of a noun.

5.7. Conclusion

This study has demonstrated that neural networks are able to flexibly learn mappings of gradable adjectives onto unmodified nominal concepts. Our neural network implementation quantitatively performs similar to previous implementations, and also provides a window into the acquisition of this type of compositional semantic structures. We found that early predictions were organized by adjective representations, with influence of the noun appearing later. We also found that the model is slowest to learn appropriate brightness predictions for concepts where the adjective and noun have contradictory brightness associations. These findings provide further insight into the mechanisms by which conceptual combination may occur and allow for more targeted hypothesis generation for future studies into the phenomena.

Chapter 6

Discussion

In this thesis, I have addressed the role of context and its influences on the nature of semantic representations and mechanisms in humans and machines. Context has been theorised to be a crucial property for structuring the semantic system (Yee & Thompson-Schill, 2016), however much of the research on language comprehension has only focused on context at the sentence-level. In this thesis, I have explored context effects at both wider (i.e., discourse) and narrower (i.e., phrasal) scales in order to better our understanding of how semantic representations and mechanisms change as a result of this context. To this end, I have presented two branches of research that explore contextual influences at these scales. In the first, I asked whether the coherence of a discourse impacts how comprehenders make downstream predictions about upcoming information. In the second, I explored how context influences the representations of concepts, both individually and in combination, analysing the extent to which perceptual feature information can be represented and flexibly combined in language models.

In **Chapter 2**, I presented three online, self-paced reading experiments that investigated whether discourse coherence is an additional cue that comprehenders use to modulate lexical predictions about upcoming material. I created three-sentence passages manipulated on coherence and predictability. Discourse coherence was manipulated by varying how consistent the information in the final target sentence was with the information presented in the two-sentence preamble contexts, while predictability was manipulated by varying the cloze probability of the critical word within the target sentence. The first experiment had a 2x2 factorial design with these stimuli and showed that participants are faster to read target sentences that are preceded by a highly coherent context, with this facilitation observed throughout the target sentence. In the second experiment, I altered the ratio of high and low coherence conditions to include a 75:25 split as previous research has suggested that comprehenders flexibly engage in predictive processing based on the reliability of the predictive cue (Brothers et al., 2017). Here, I found a similar result as Experiment 1. In the third and final experiment, I replaced the low predictability condition, which was created by substituting the most predictable critical word with a lexical item that was less plausible but not impossible, with a semantically anomalous condition. I did this as an attempt to strengthen the predictability manipulation. The results showed that participants are faster to read highly predictable critical words than anomalous ones; and that, anomalous critical words slowed reading to a greater extent when preceded by highly coherent contexts. In sum, I found that comprehenders are sensitive to subtle topic shifts in discourse and that they will use cues at both global and local contexts but will do so especially when there are violations present at the local sentence level.

Chapter 3 introduced modelling work that takes the RT dataset collected in Chapter 2 and investigates whether surprisal values from Large Language Models reflects the type of longer context effects I observed in humans. Across two sets of

analyses, I asked whether next-word predictions from GPT-2 are significantly influenced by the same linguistic features that cause significant differences in human RTs, and whether surprisal explains additional variance in predicting these RTs. I found evidence that GPT-2 is sensitive to topic shifts within a passage with better model fits for all models that included the experimental conditions. For these linear mixed effects models that included coherence and predictability as predictors, I found a coherence effect at the pre-critical ROI, which did not extend into the post-critical ROI; and I found observed a graded predictability effect at the post-critical ROI, where I saw the lowest surprisal for trials containing a highly predictable critical word, mid-level surprisal values for the trials with a less predictable critical word, and the highest surprisal for trials containing an anomalous critical word. Moreover, for my RT analyses, I found that surprisal does account for some additional variance when predicting RTs. This was reflected in robust large effects of surprisal across the target sentence for each experiment. However, it does not fully account for the patterns observed in the RT dataset, with the effects of coherence observed in Chapter 2 still remaining. This suggests that the benefit of a highly coherent context is not explained by simply lowering linguistic surprisal, which opens up the possibility that comprehenders are making use of discourse-related models to facilitate their comprehension of such narratives.

In **Chapter 4**, I presented an investigation into how well word embeddings from Distributional Semantic Models and Large Language Models can be used to predict human judgements about the perceptual features of words. In particular, I focused on the perceptual feature of brightness as previous research has demonstrated that it is not well represented in word embeddings. I also included the perceptual feature of shape as this has been shown to be well represented (Chersoni et al., 2021; Turton et al., 2020; Utsumi, 2020). Across two experiments, I explored whether the presence of additional context would improve performance on the perceptual prediction task by making use of Transformer-based LLMs' ability to represent word meaning in context. In the first experiment, I focused on representations of nouns and compared decontextualised and contextualised Word2Vec and BERT embeddings for a large number of concepts. I found modest prediction of brightness, suggesting that even difficult perceptual properties can be predicted when the dataset includes concepts where the feature is particularly salient. In addition, I found very good prediction of shape, replicating previous results (Chersoni et al., 2021; Turton et al., 2020; Utsumi, 2020). In my evaluation of my embeddings, context-free BERT embeddings outperformed Word2Vec embeddings, adding to the literature that an aggregated contexutalised representation is a better predictor of semantic feature ratings than static embeddings (Bommasani et al., 2020; Turton et al., 2021). However, the addition of context had no influence. In the second experiment, I focused on representations of adjective–noun phrases and only predicted the brightness of concepts. Overall, I found good prediction of brightness, with the embeddings mimicking the non-additive effect demonstrated in the human data. Here, context had a limited impact on prediction performance, such that including a prompt towards the target feature had a slight influence on performance. Together, these results demonstrate that some amount of perceptual feature information can be learnt from linguistic input alone, and that strong

embodiment is not necessary to explain how perceptual properties may be represented in semantic systems.

Finally, **Chapter 5** details my investigations into semantic flexibility within conceptual combinations. I explored how a neural network encodes the semantic flexibility inherent in gradable adjectives. I again focused on adjective–noun phrases, investigating the contextual influence of gradable adjectives on phrasal-level representations, using the perceptual feature of brightness. I found that the flexible learning of gradable adjectives was possible, with predictions first made based on the adjective alone, and then further modulated by noun information later in learning. I also discovered that the model outputs mimicked the type of non-additive feature modulation present in the human data used.

6.1. Theoretical implications

Overall, my investigations into the influences of context at disparate scales other than the sentence-level demonstrate the importance of context for a unified theory of semantic comprehension. Concerning wider contextual influences, I have demonstrated that comprehenders are sensitive to discourse-level cues when forming expectations about upcoming material and that these types of facilitative effects are not entirely explained by lowering linguistic surprisal. Meanwhile, for phrasal-level contextual influences, I have also shown that neural networks can encode the semantic flexibility of gradable adjectives, mimicking the non-additive feature modulation observed in human judgements. Furthermore, I have explored the representation of perceptual information in pre-trained word embeddings from both DSMs and LLMs.

6.1.1. Predictive processing and surprisal theory

Within the predictive processing literature, a number of linguistic properties have already been studied with regards to their influence on upcoming expectations. These properties include the grammatical gender and the definiteness of a referent, among others (Carter & Nieuwland, 2022; Fleur et al., 2020; Wicha et al., 2003, 2004). Evidence from Chapter 2 furthers our understanding of the linguistic properties that comprehenders are sensitive to by demonstrating that discourse coherence is another such cue. This finding also adds to our understanding of how the global discourse impacts local predictive processing. The results presented are consistent with prior evidence that the presence of a highly coherent discourse context facilitates comprehension, while discrepancies with the existing discourse model results in a slowdown (Hess et al., 1995; Myers & O'Brien, 1998; O'Brien & Albrecht, 1992; Stewart et al., 2009).

Theoretically, both the results from Chapters 2 and 3 are in accord with the idea that comprehenders build a discourse model of the events portrayed within a narrative (Kintsch, 1988; Zwaan, 2016; Zwaan & Radvansky, 1998). Of note, this finding does not necessarily adjudicate between theoretical perspectives as to the cause of facilitation effects in language processing. It is possible that the presence of a discourse-based

situation model could facilitate processing in the ways specified by both the integrationist and anticipatory predictive processing accounts. As such, discussion of these results is largely framed as theory-agnostic, leaving future work to touch on these important issues. Within Chapter 3, I found that the effects of coherence remained even with the addition of surprisal as a predictor for the RT analyses. I take this as evidence that the benefit of a highly coherent context is not simply due to lowering linguistic surprisal, but rather that an additional level of processing could be occurring associated with the presence of a discourse model. However, these results do not align with recent work from Michaelov et al. (2023), who find that LM surprisal can account for discourse-level effects in the N400 data collected by Nieuwland and van Berkum (2006). In the chapter, I present discussion of the differences between my work and Michaelov and colleagues which could explain this difference, namely methodological differences between SPR and ERP data, and the fact that the discourse-level cue analysed is different between the two studies. However, recent research by Lopopolo and Rabovsky (2024) demonstrated that both surprisal and Semantic Update (SU) from their Sentence Gestalt model had a significant effect on N400 amplitudes (taken from Frank et al., 2015). Taken together, these findings suggest that surprisal may not capture all aspects of incremental language processing.

6.1.2. Conceptual processing and embodied cognition

The work presented in Chapter 4 relates to some of the predictions from embodied cognition. Overall, the results align with previous research suggesting that a surprising amount of perceptual information can be transferred through linguistic content alone (Chersoni et al., 2021; Utsumi, 2020). Utsumi (2020) argues that the type of statistics used by language models implicitly contain conceptual knowledge from direct experiences through the way in which language is used to communicate about experiences. The finding that the perceptual feature of brightness can be predicted to some extent, but only when it is a salient feature of the concepts included also indicates that performance may be dependent on the set of concepts included. Generally, my results suggest that a strong embodiment account is not necessary for building an adequate semantic system. Instead, it has been theorised that the semantic information which comprehenders take in through perception and action and the semantic information which is transferred through language supplement one another to form a semantically rich representation of the world (Louwerse & Connell, 2011; Louwerse & Jeuniaux, 2010).

The findings from Chapter 5 give insight into a number of aspects related to the acquisition of concepts and the compositional processes that occur during conceptual combination. I demonstrated that earlier in training, predictions were made based solely on the adjective, while during later stages of training, information from the noun was used to modulate predictions. This presents one possible hypothesis for how adjective–noun phrases that contain a gradable adjective may be composed (Barsalou, 2017). Moreover, I observed that the hidden representations of the model were largely clustered by adjective, reinforcing the idea that the adjective is the dominant organizing principle for the model's representations. When comparing my implementation with those detailed in Solomon and Thompson-Schill (2020), I found that my neural network

performed as well as their Bayesian model but was trained on a smaller dataset of averaged ratings. Whereas their Bayesian model was trained on the distribution of ratings across individual participants, with the authors arguing that a construct of feature uncertainty was crucial for accurate modelling of this semantic flexibility. In contrast, my implementation does not include a concept of feature uncertainty, which I take as evidence that feature uncertainty is not necessarily required to explain the semantic flexibility in conceptual combinations.

6.1.3. Cognitive plausibility of models

The findings from Chapters 3, 4 and 5 have relevance for our understanding on the cognitive plausibility of computational models. In Chapters 3 and 4, I use pre-trained language models (GPT-2, Word2Vec and BERT) across my investigations, touching on the applicability and limits of using pre-trained language models to answer cognitively motivated questions. For example, in Chapter 4, I evaluate the representation of perceptual information in word embeddings using a dataset of human ratings. In a similar vein, Utsumi (2020) emphasises the importance of knowing the degree of representational ability of word embeddings used within the cognitive sciences in order to improve performance within cognitive modelling and practical NLP tasks. The issue of cognitive plausibility with regards to models is not a trivial question that can be answered simply. There is acceptance that no model will provide a full account of psychological semantics in the near future (Lake & Murphy, 2023). However, as eloquently stated by Lake and Murphy (2023), there is merit in asking whether a model exhibits similar processes to a human; and if this is not the case, presenting an analysis as to the differences between model outputs and those collected from humans. Chapter 5, in particular, is an example of this. In this chapter, I present an investigation into how a neural network encodes the semantic flexibility inherent in gradable adjectives and the nouns that they modify. Within this, I include analysis of a subset of concepts, identifying which concepts the model learns first and the incidence of errors. Overall, the work presented in my thesis makes use of computational models as an additional tool to study the influence of context on semantic representations and mechanisms. This work adds to previous research using computational models to study aspects of human language processing that would otherwise be impossible to test empirically (Lake & Murphy, 2023; Michaelov et al., 2021).

6.2. Limitations and future directions

Chapter 2 introduced work that explored the factors influencing engagement in predictive processing. A clear limitation of the presented work lies in the methods used. In this work, I used online self-paced reading (SPR) experiments to evaluate the impact of coherence on sentence processing. However, self-paced reading is not as temporally accurate as other methods used to study predictive processing, such as eye-tracking and EEG. My reason for using online SPR experiments was due to lab closure and inaccessibility at the time, however future work should investigate if these findings generalise across experimental methods. Indeed, using neuroimaging (namely, EEG) to study discourse coherence relations would allow for a more temporally accurate

assessment of the predictive mechanisms at play when comprehenders are faced with subtle topic shifts. Moreover, one of the theoretical motivations for this work was to investigate how the maintenance and updating of a situation model impacts predictive processing. Indeed, one current theory is that comprehenders generate probabilistic predictions at multiple levels of linguistic representation, in the form of hierarchical generative models (Brothers et al., 2020; Kuperberg, 2021; Kuperberg et al., 2020). Future work that investigates the interactions between predictions made at different levels of representation, for example at the discourse-level, could further our understanding of the environmental and linguistic factors that influence how predictions are formed (Brothers et al., 2019; Dave et al., 2021).

The work in Chapter 3 theoretically follows on from the questions asked in Chapter 2, however, the presented work includes specific choices with regards to the computational models used. For example, I chose one pre-trained LLM (GPT-2) to explore the relationship between surprisal and my reaction time data. This can be thought of as a limitation of the current work, because as it stands, it is not clear if these results generalise across models. Theoretically, this limitation touches on a current debate within the computational psycholinguistic literature that questions why surprisal from cognitively implausible models, such as Transformer-based models, can be used to predict human reaction times (Frank et al., 2019; Merkx & Frank, 2021; Michaelov et al., 2021). Future work that studies the computational factors influencing how well surprisal can predict measures of human language processing on datasets that manipulate global discourse cues would broaden our understanding of both predictive language models and predictive processing. For example, comparing surprisal from different model architectures or including models that have been pre-trained on size-limited corpora in an effort to match the amount of linguistic input a child may receive could be informative and would add to a growing body of research (Michaelov & Bergen, 2022; Warstadt et al., 2023; Wilcox et al., 2020).

In Chapter 4, I presented work exploring the representational ability of word embeddings. This work touches on a key debate in the cognitive sciences, that is whether language processing is symbolic or embodied (Andrews et al., 2014; Dove, 2022; Louwerse, 2018). In addition, the nature of the representations themselves, as well as the extent to which perceptual information is crucial to their makeup, is also under debate (Kiefer & Pulvermüller, 2012). Future work aimed at understanding the role and significance of perceptual information in word embeddings could be a fruitful research direction. In particular, investigating how word embeddings represent multiple types of perceptual information, other than brightness and shape information, and if these are sensitive to context could be an interesting direction for further research. In Chapter 5, I focused on the semantic flexibility present in adjective–noun phrases. Current understanding of the idiosyncratic nature of the compositional process is limited and represents a clear gap in current knowledge across the cognitive sciences (Coutanche et al., 2019). Further study of this research area, such as experiments on modification with different types of adjectives (e.g., absolute gradable adjectives), could extend our knowledge on compositional processes in language in general.

6.3. Conclusion

In sum, this thesis has explored how context influences semantic representations and mechanisms in humans and machines. Through this, I have shown that understanding context is vital in order to have a full picture of the semantic system. In my work, I defined context on both wider and narrower scales than the sentence level, however there is no clear definition of what context can encompass. As such, further investigations into the types of linguistic cues which impact semantic representations and the mechanisms that comprehenders engage in during processing will help refine the meaning of context. A true understanding of what humans are doing when processing language is still a long way off, but this thesis is one small attempt to advance what we know about meaning.

Appendix

Combined Experiment 1 and 3 Preambles

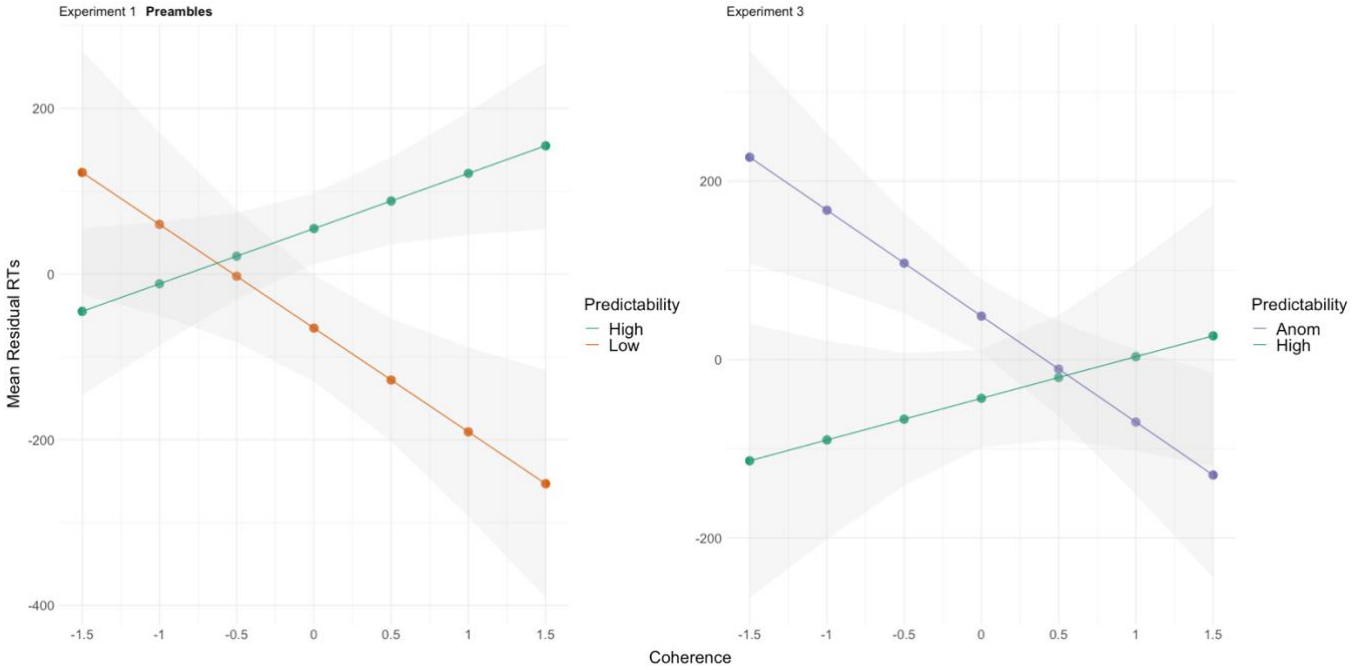

Figure 7.1. Results for preamble region for Experiment 1 (left) and Experiment 3 (right); shaded areas denote standard error.

Pre-registered analyses on initial batch of participants

Here we report the results of the pre-registered analyses with our first batch of 52 participants. Participants were excluded using the same criteria as reported in the main text for the pre-registered analyses of Experiment 1 (n= 21). As such, our final pool for these analyses was n=31.

Preambles. Participants had similar reading times for high and less coherent preamble contexts (β= -0.003, SE= 0.01, *t*= 0.23, *p*= 0.81), as well as for high and less predictable critical words (β= -0.0004, SE= 0.01, *t*= -0.04, *p*= 0.97), with no interaction (β= 0.001, SE= 0.009, *t*= 0.10, *p*= 0.92). These results confirm that participants took a similar time to read the preambles across all conditions.

Pre-critical Word. For the pre-critical ROI, we found that participants had similar reading times in this region for high and less coherent trials (β= -0.005, SE= 0.003, *t*= -1.59, *p*= 0.11). There was no difference in reading times between high and low predictable conditions (β= -0.0007, SE= 0.003, *t*= -0.22, *p*= 0.83), which was expected as this region precedes the critical word. Further, no interaction effect was observed (β= -0.002, SE= 0.003, *t*= -0.74, *p*= 0.46).

Critical Word and Spillover. At the critical word and spillover region, our main ROI, there was a marginal difference in reading times for highly predictable critical words compared to less predictable (β= -0.01, SE= 0.007, *t*= -1.73, *p*= 0.09). We also found significant effects of critical word length (β= 0.01, SE= 0.005, *t*= 2.42, *p*= 0.02) and frequency (β= -0.01, SE= 0.005, *t*= -2.07, *p*= 0.04). We found no processing advantage for critical words preceded by high coherent conditions, compared to less coherent (β= 0.001, SE= 0.005, *t*= 0.29, *p*= 0.77), and no interaction effect (β= 0.001, SE= 0.005, *t*= 0.22, *p*= 0.83).

Critical Word. At the critical word, there was a marginal effect of predictability, with faster reading times for highly predictable critical words compared to less predictable (β= -0.01, SE= 0.007, *t*= -1.71, *p*= 0.09). We also found significant differences for critical word length (β= 0.01, SE= 0.005, *t*= 2.42, *p*= 0.02) and frequency (β= -0.01, SE= 0.005, *t*= -2.08, *p*= 0.04). There was no effect of coherence, with similar reading times for critical words preceded by high and less coherent preambles (β= 0.002, SE= 0.005, *t*= 0.35, *p*= 0.72), and no interaction effect (β= 0.001, SE= 0.005, *t*= 0.22, *p*= 0.83).

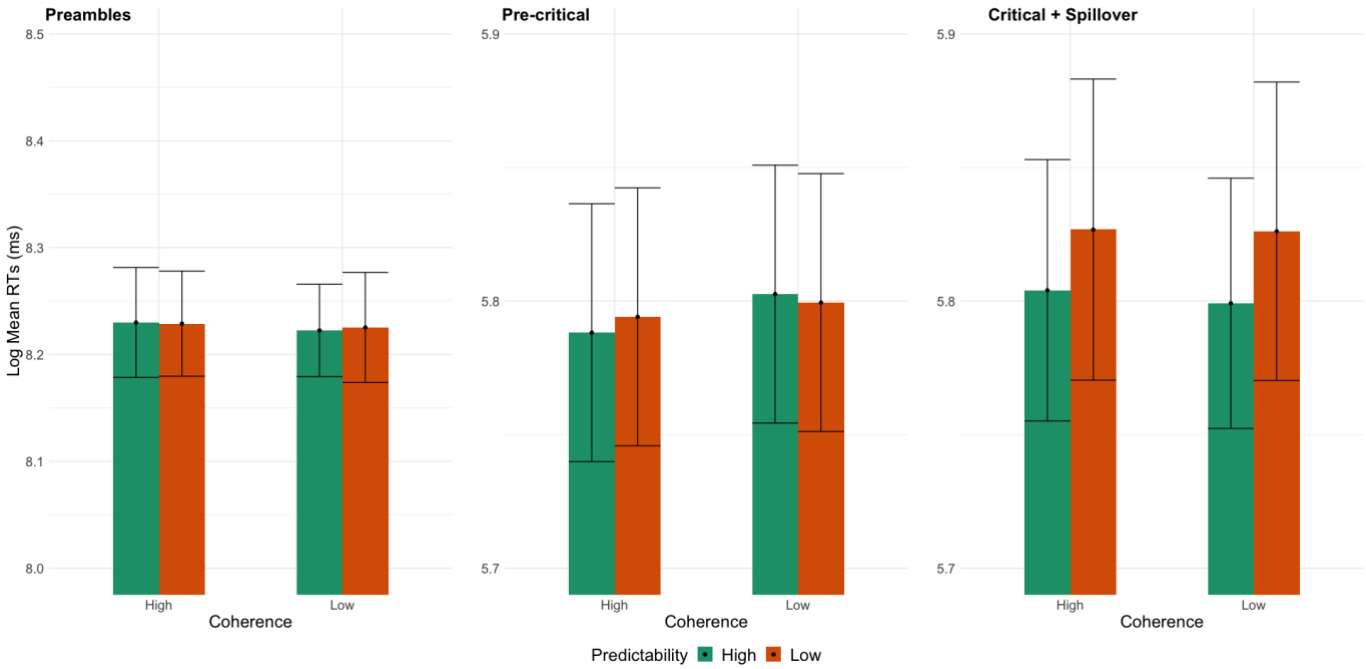

Figure 7.2. Experiment 1 pre-registered analysis results from initial batch of 52 participants; error bars indicate the 95% confidence intervals.

Table 7.1. Standard deviation across MSE scores for each run (10) of model on Solomon noun dataset.

	MSE	Standard deviation (σ)
Onehot	0.09	0.002
Word2Vec	0.08	0.002
BERTbase colour-contextually prompted	0.06	0.002
BERTbase brightness-contextually prompted	0.07	0.003
BERTbase context-free	0.06	0.002
BERTLarge colour-contextually prompted	0.03	0.001
BERTLarge brightness-contextually prompted	0.04	0.002
BERTLarge context-free	0.04	0.001

Table 7.2. Standard deviation across MSE scores for each run (10) of model on Binder dataset predicting brightness.

	MSE	Standard deviation (σ)
Onehot		
Concrete	0.01	8.71520102225854e-05
Concrete+abstract	0.01	4.3093925851674064e-05
Word2Vec		
Concrete	0.01	0.0004
Concrete+abstract	0.01	0.0002
BERTbase colour-contextually prompted		
Concrete	0.01	0.0004
Concrete+abstract	0.01	0.0004
BERTbase brightness-contextually prompted		
Concrete	0.02	0.0003
Concrete+abstract	0.01	0.0003
BERTbase context-free		
Concrete	0.02	0.0002
Concrete+abstract	0.01	0.0002
BERTLarge colour-contextually prompted		
Concrete	0.02	0.0002
Concrete+abstract	0.02	0.0002
BERTLarge brightness-contextually prompted		
Concrete	0.02	0.0003
Concrete+abstract	0.02	0.0003
BERTLarge context-free		
Concrete	0.02	0.0002
Concrete+abstract	0.01	0.0002

Table 7.3. Standard deviation across MSE scores for each run (10) of model on Binder dataset predicting shape.

	MSE	Standard deviation (σ)
Onehot		
Concrete	0.04	0.0002
Concrete+abstract	0.11	0.0007
Word2Vec		
Concrete	0.02	0.0003
Concrete+abstract	0.02	0.0003
BERTbase contextually prompted		
Concrete	0.01	0.0004
Concrete+abstract	0.02	0.0003
BERTbase context-free		
Concrete	0.01	0.0002
Concrete+abstract	0.02	0.0004
BERTLarge contextually prompted		
Concrete	0.02	0.0004
Concrete+abstract	0.06	0.002
BERTLarge context-free		
Concrete	0.01	0.0001
Concrete+abstract	0.02	0.0003

Table 7.4. Standard deviation across MSE scores for each run (10) of model on Solomon adjective-noun dataset.

	MSE	Standard deviation (σ)
Onehot		
Dark	0.10	0.01
Light	0.08	0.01
Word2Vec		
Dark	0.01	0.006
Light	0.01	0.002
BERTbase colour-contextually prompted		
Dark	0.01	0.0007
Light	0.01	0.0009
BERTbase brightness-contextually prompted		
Dark	0.01	0.0005
Light	0.01	0.001
BERTbase context-free		
Dark	0.01	0.003
Light	0.02	0.007
BERTLarge colour-contextually prompted		
Dark	0.02	0.001
Light	0.01	0.001
BERTLarge brightness-contextually prompted		
Dark	0.01	0.001
Light	0.02	0.0004
BERTLarge context-free		
Dark	0.01	0.004
Light	0.02	0.005

Bibliography

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 109–132. https://doi.org/10.18653/v1/2021.conll-1.9

Abnar, S., Ahmed, R., Mijnheer, M., & Zuidema, W. (2018). Experiential, Distributional and Dependency-based Word Embeddings have Complementary Roles in Decoding Brain Activity. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 57–66. https://doi.org/10.18653/v1/W18-0107

Agarap, A. F. (2019). *Deep Learning using Rectified Linear Units (ReLU)* (arXiv:1803.08375). arXiv. https://doi.org/10.48550/arXiv.1803.08375

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a Mental Model: Maintaining Both Local and Global Coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1061–1070. https://psycnet.apa.org/buy/1994-00399-001

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science*, *33*(4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Amari, S. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, *5*(4), 185–196. https://doi.org/10.1016/0925-2312(93)90006-O

Amsel, B. D., DeLong, K. A., & Kutas, M. (2015). Close, but no garlic: Perceptuomotor and event knowledge activation during language comprehension. *Journal of Memory and Language*, *82*, 118–132. https://doi.org/10.1016/j.jml.2015.03.009

Andrews, M., Frank, S., & Vigliocco, G. (2014). Reconciling Embodied and Distributional Accounts of Meaning in Language. *Topics in Cognitive Science*, *6*(3), 359–370. https://doi.org/10.1111/tops.12096

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*(4), 1407–1425. https://doi.org/10.3758/s13428-020-01501-5

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Asher, N. (2011). *Lexical Meaning in Context: A Web of Words*. Cambridge University Press.

Aurnhammer, C., & Frank, S. L. (2018). *Comparing gated and simple recurrent neural network architectures as models of human sentence processing* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/wec74

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarrell, N. S., & Nitsch, K. (1974). Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior*, *13*(4), 471–481. https://doi.org/10.1016/S0022-5371(74)80024-1

Barner, D., & Snedeker, J. (2008). Compositionality and Statistics in Adjective Acquisition: 4-Year-Olds Interpret Tall and Short Based on the Size Distributions of Novel Noun Referents. *Child Development*, *79*(3), 594–608. https://doi.org/10.1111/j.1467-8624.2008.01145.x

Baroni, M. (2013). Composition in Distributional Semantics. *Language and Linguistics Compass*, *7*(10), 511–522. https://doi.org/10.1111/lnc3.12050

Baroni, M., & Zamparelli, R. (2010). Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1183–1193. https://aclanthology.org/D10-1115.pdf

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition*, *10*(1), 82–93. https://doi.org/10.3758/BF03197629

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227. https://doi.org/10.3758/BF03196968

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660. https://doi.org/10.1017/S0140525X99002149

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2010). Grounded Cognition: Past, Present, and Future. *Topics in Cognitive Science*, *2*(4), 716–724. https://doi.org/10.1111/j.1756-8765.2010.01115.x

Barsalou, L. W. (2017). Cognitively Plausible Theories of Concept Composition. In J. A. Hampton & Y. Winter (Eds.), *Compositionality and Concepts in Linguistics and Psychology* (Vol. 3, pp. 9–30). Springer Open. https://link.springer.com/chapter/10.1007/978-3-319-45977-6_2

Barsalou, L. W., Dutriaux, L., & Scheepers, C. (2018). Moving beyond the distinction between concrete and abstract concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170144. https://doi.org/10.1098/rstb.2017.0144

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and Embodiment: Debates on meaning and cognition* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199217274.003.0013

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*(2), 84–91. https://doi.org/10.1016/S1364-6613(02)00029-3

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *arXiv:1406.5823*. http://arxiv.org/abs/1406.5823

Bemis, D. K., & Pylkkanen, L. (2011). Simple Composition: A Magnetoencephalography Investigation into the Comprehension of Minimal Linguistic Phrases. *Journal of Neuroscience*, *31*(8), 2801–2814. https://doi.org/10.1523/JNEUROSCI.5003-10.2011

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Berends, S. M., Brouwer, S. M., & Sprenger, S. A. (2016). Eye-Tracking and the Visual World Paradigm. In M. S. Schmid, S. M. Berends, C. Bergmann, S. M. Brouwer, N.

Meulman, B. J. Seton, S. A. Sprenger, & L. A. Stowe (Eds.), *Designing Research on Bilingual Development: Behavioral and Neurolinguistic Experiments* (pp. 55–80). Springer International Publishing. https://doi.org/10.1007/978-3-319-11529-0_5

Berkum, J. J. A. van, Hagoort, P., & Brown, C. M. (1999). Semantic Integration in Sentences and Discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*, *11*(6), 657–671. https://doi.org/10.1162/089892999563724

Bermeitinger, C., Wentura, D., & Frings, C. (2011). How to switch on and switch off semantic priming effects for natural and artifactual categories: Activation processes in category memory depend on focusing specific feature dimensions. *Psychonomic Bulletin & Review*, *18*(3), 579–585. https://doi.org/10.3758/s13423-011-0067-z

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, *25*(10), 883–895. https://doi.org/10.1016/j.tics.2021.07.006

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3–4), 130–174. https://doi.org/10.1080/02643294.2016.1147426

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience Grounds Language. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8718–8735). https://doi.org/10.18653/v1/2020.emnlp-main.703

Black, A. (1988). The syntax of conversational coherence. *Discourse Processes*, *11*(4), 433–455. https://doi.org/10.1080/01638538809544712

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, *6*, 213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Boleda, G., Baroni, M., Pham, T. N., & McNally, L. (2013). Intensionality was only alleged: On adjective-noun composition in distributional semantics. *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, 35–46. https://aclanthology.org/W13-0104

Bolognesi, M., & Steen, G. (2018). Editors' Introduction: Abstract Concepts: Structure, Processing, and Modeling. *Topics in Cognitive Science*, *10*(3), 490–500. https://doi.org/10.1111/tops.12354

Bommasani, R., Davis, K., & Cardie, C. (2020). Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4758–4781. https://doi.org/10.18653/v1/2020.acl-main.431

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of Life Reviews*, *29*, 120–153. https://doi.org/10.1016/j.plrev.2018.12.001

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language

comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 607–624. https://doi.org/10.3758/s13415-015-0340-0

Brothers, T., Dave, S., Hoversten, L. J., Traxler, M. J., & Swaab, T. Y. (2019). Flexible predictions during listening comprehension: Speaker reliability affects anticipatory processes. *Neuropsychologia*, *135*, 107225. https://doi.org/10.1016/j.neuropsychologia.2019.107225

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 104174. https://doi.org/10.1016/j.jml.2020.104174

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, *93*, 203–216. https://doi.org/10.1016/j.jml.2016.10.002

Brothers, T., Wlotko, E. W., Warnke, L., & Kuperberg, G. R. (2020). Going the Extra Mile: Effects of Discourse Context on Two Late Positivities During Language Comprehension. *Neurobiology of Language*, *1*(1), 135–160. https://doi.org/10.1162/nol_a_00006

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526. https://doi.org/10.3758/BF03193020

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain and Language*, *98*(2), 159–168. https://doi.org/10.1016/j.bandl.2006.04.005

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, *56*(1), 103–128. https://doi.org/10.1016/j.jml.2006.07.005

Carter, G.-A., & Nieuwland, M. S. (2022). Predicting Definite and Indefinite Referents During Discourse Comprehension: Evidence from Event-Related Potentials. *Cognitive Science*, *46*(2). https://doi.org/10.1111/cogs.13092

Casasanto, D., & Lupyan, G. (2015). All Concepts Are Ad Hoc Concepts. In E. Margolis & S. Laurence (Eds.), *The Conceptual Mind* (pp. 543–566). The MIT Press. https://doi.org/10.7551/mitpress/9383.003.0031

Chatterjee, A. (2010). Disembodying cognition. *Language and Cognition*, *2*(1), 79–116. https://doi.org/10.1515/langcog.2010.004

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). *One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling* (arXiv:1312.3005). arXiv. http://arxiv.org/abs/1312.3005

Chersoni, E., Santus, E., Huang, C.-R., & Lenci, A. (2021). Decoding Word Embeddings with Brain-Based Semantic Features. *Computational Linguistics*, *47*(3), 663–698. https://doi.org/10.1162/coli_a_00412

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators* (arXiv:2003.10555). arXiv. https://doi.org/10.48550/arXiv.2003.10555

Cohen, B., & Murphy, G. L. (1984). Models of concepts. *Cognitive Science*, *8*(1), 27–58. https://doi.org/10.1016/S0364-0213(84)80024-5

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1

Connell, L., & Lynott, D. (2014). Principles of Representation: Why You Can't Represent the Same Concept Twice. *Topics in Cognitive Science*, *6*(3), 390–406. https://doi.org/10.1111/tops.12097

Costello, F. J., & Keane, M. T. (2000). Efficient Creativity: Constraint-Guided Conceptual Combination. *Cognitive Science*, *24*(2), 299–349. https://doi.org/10.1207/s15516709cog2402_4

Coutanche, M. N., Solomon, S., & Thompson-Schill, S. L. (2019). Conceptual Combination in The Cognitive Neurosciences. In D. Poeppel, G. R. Mangun, & M. S. Gazzaniga (Eds.), *The Cognitive Neurosciences* (6th ed.). MIT Press. https://doi.org/10.31234/osf.io/9jptv

Dave, S., Brothers, T., Hoversten, L. J., Traxler, M. J., & Swaab, T. Y. (2021). Cognitive control mediates age-related changes in flexible anticipatory processing during listening comprehension. *Brain Research*, *1768*, 147573. https://doi.org/10.1016/j.brainres.2021.147573

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394. https://doi.org/10.1098/rstb.2012.0394

Delogu, F., Crocker, M. W., & Drenhaus, H. (2017). Teasing apart coercion and surprisal: Evidence from eye-movements and ERPs. *Cognition*, *161*, 46–59. https://doi.org/10.1016/j.cognition.2016.12.017

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162. https://doi.org/10.1016/j.neuropsychologia.2014.06.016

Delong, K. A., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP responses to low cloze probability sentence continuations. *Psychophysiology*, *48*(9), 1203–1207. https://doi.org/10.1111/j.1469-8986.2011.01199.x

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

Demonte, V. (2019). Adjectives. In C. Maienborn, K. Von Heusinger, & P. Portner (Eds.), *Semantics—Lexical Structures and Adjectives*. De Gruyter Mouton. https://doi.org/10.1515/9783110626391

Derby, S., Miller, P., & Devereux, B. (2019). Feature2Vec: Distributional semantic modelling of human property knowledge. In K. Inui, J. Jiang, V. Ng, & X. Wan

(Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5853–5859). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1595

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, *46*(4), 1119–1127. https://doi.org/10.3758/s13428-013-0420-4

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Dixon, R. M. W., & Aikhenvald, A. Y. (2004). *Adjective Classes: A Cross-Linguistic Typology*. OUP Oxford.

Dove, G. O. (2014). Thinking in Words: Language as an Embodied Medium of Thought. *Topics in Cognitive Science*, *6*(3), 371–389. https://doi.org/10.1111/tops.12102

Dove, G. O. (2016). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychonomic Bulletin & Review*, *23*(4), 1109–1121. https://doi.org/10.3758/s13423-015-0825-4

Dove, G. O. (2022). Rethinking the role of language in embodied cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1870), 20210375. https://doi.org/10.1098/rstb.2021.0375

Ellis, C., Henderson, A., Wright, H. H., & Rogalski, Y. (2016). Global coherence during discourse production in adults: A review of the literature: Global coherence during discourse production in adults: a review. *International Journal of*

Language & Communication Disorders, *51*(4), 359–367. https://doi.org/10.1111/1460-6984.12213

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Enochson, K., & Culbertson, J. (2015). Collecting Psycholinguistic Response Time Data Using Amazon Mechanical Turk. *PLOS ONE*, *10*(3), e0116946. https://doi.org/10.1371/journal.pone.0116946

Estes, Z. (2003). Attributive and relational processes in nominal combination. *Journal of Memory and Language*, *48*(2), 304–319. https://doi.org/10.1016/S0749-596X(02)00507-7

Estes, Z., & Jones, L. L. (2006). Priming via relational similarity: A copper horse is faster when seen through a glass eye. *Journal of Memory and Language*, *55*(1), 89–101. https://doi.org/10.1016/j.jml.2006.01.004

Ettinger, A. (2020). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. https://doi.org/10.1162/tacl_a_00298

Făgărăşan, L., Vecchi, E. M., & Clark, S. (2015). From distributional semantics to feature norms: Grounding semantic models in human perceptual data. *Proceedings of the 11th International Conference on Computational Semantics*, 52–57. https://aclanthology.org/W15-0107.pdf

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84. https://doi.org/10.1016/j.brainres.2006.06.101

Ferreira, F., & Clifton, C. (1986). The Independence of Syntactic Processing. *Journal of Memory and Language*, *25*(3), 348–368. https://www.proquest.com/docview/1297341868/citation/C8798D31BDDB485EPQ/1

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in linguistic analysis* (pp. 10–32). Blackwell.

Fleur, D. S., Flecken, M., Rommers, J., & Nieuwland, M. S. (2020). Definitely saw it coming? The dual nature of the pre-nominal prediction effect. *Cognition*, *204*, 104335. https://doi.org/10.1016/j.cognition.2020.104335

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Forster, K. I. (1979). Levels of processing and the structure of the language processor. *Sentence Processing*, 27–85. https://cir.nii.ac.jp/crid/1572543024356836224

Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 465–495. https://doi.org/10.1080/14640748108400804

Forster, K., I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680.

Frank, S. L. (2013). Uncertainty Reduction as a Measure of Cognitive Load in Sentence Comprehension. *Topics in Cognitive Science*, *5*(3), 475–494. https://doi.org/10.1111/tops.12025

Frank, S. L., Monaghan, P., & Tsoukala, C. (2019). Neural Network Models of Language Acquisition and Processing. In P. Hagoort (Ed.), *Human Language* (pp. 277–292). The MIT Press. https://doi.org/10.7551/mitpress/10841.003.0026

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 878–883. https://repository.ubn.ru.nl/bitstream/handle/2066/119221/119221.pdf;

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203. https://doi.org/10.1080/23273798.2017.1323109

French, R. M., & Labiouse, C. (2002). Four problems with extracting human semantics from large text corpora. *Proceedings of the Twenty- Fourth Annual Conference of the Cognitive Science Society*, *24*, 316–322. https://escholarship.org/uc/item/4c51156n

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), 137–146. https://doi.org/10.1007/s11222-009-9153-8

Futrell, R., Gibson, E., Tily, H., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2017). *The Natural Stories Corpus* (arXiv:1708.05763). arXiv. https://doi.org/10.48550/arXiv.1708.05763

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 688–698. https://doi.org/10.18653/v1/E17-1065

Gagné, C. L. (2001). Relation and lexical priming during the interpretation of noun–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 236–254. https://doi.org/10.1037/0278-7393.27.1.236

Galetzka, C. (2017). The Story So Far: How Embodied Cognition Advances Our Understanding of Meaning-Making. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01315

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, *70*(350), 320–328. https://doi.org/10.1080/01621459.1975.10479865

Gibbs Jr, R. W. (2005). *Embodiment and Cognitive Science* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511805844

Glenberg, A. M. (2010). Embodiment as a unifying perspective for psychology. *WIREs Cognitive Science*, *1*(4), 586–596. https://doi.org/10.1002/wcs.55

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*(3), 558–565. https://doi.org/10.3758/BF03196313

Glenberg, A. M., & Mehta, S. (2008). Constraint on covariation: It's not meaning. *Italian Journal of Linguistics*, *20*(1), 241–264.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning. *Journal of Memory and Language*, *43*(3), 379–401. https://doi.org/10.1006/jmla.2000.2714

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 10–18. https://doi.org/10.18653/v1/W18-0102

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, *6*(7), 975–987. https://doi.org/10.1038/s41562-022-01316-8

Grice, H. P. (1975). Logic and Conversation. In *Speech acts* (pp. 41–58). Brill.

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*(4), 626–653. https://doi.org/10.1080/17470218.2015.1038280

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Haeuser, K. I., & Kray, J. (2022). How odd: Diverging effects of predictability and plausibility violations on sentence reading and word memory. *Applied Psycholinguistics*, *43*(5), 1193–1220. https://doi.org/10.1017/S0142716422000364

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science*, *304*(5669), 438–441. https://doi.org/10.1126/science.1095455

Hald, L. A., Steenbeek-Planting, E. G., & Hagoort, P. (2007). The interaction of discourse context and world knowledge in online sentence comprehension. Evidence from the N400. *Brain Research*, *1146*, 210–218. https://doi.org/10.1016/j.brainres.2007.02.054

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1–8. https://doi.org/10.3115/1073336.1073357

Hale, J. (2003). The Information Conveyed by Words in Sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123. https://doi.org/10.1023/A:1022492123056

Hale, J. (2006). Uncertainty About the Rest of the Sentence. *Cognitive Science*, *30*(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64

Hale, J. (2016). Information-theoretical Complexity Metrics. *Language and Linguistics Compass*, *10*(9), 397–412. https://doi.org/10.1111/lnc3.12196

Halff, H. M., Ortony, A., & Anderson, R. C. (1976). A context-sensitive representation of word meanings. *Memory & Cognition*, *4*(4), 378–383. https://doi.org/10.3758/BF03213193

Hampton, J. A. (2015). Categories, prototypes and exemplars. In *The Routledge Handbook of Semantics* (1st ed.). Routledge. https://doi.org/10.4324/9781315685533

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Harris, Z. (1954). Distributional structure. *Word*, *10*, 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hartung, M., Kaupmann, F., Jebbara, S., & Cimiano, P. (2017). Learning Compositionality Functions on Word Embeddings for Modelling Attribute Meaning in Adjective-Noun Phrases. *Proceedings of the 15th Conference of the European Chapter of Association for Computational Linguistics: Volume 1, Long Papers*, 54–64. https://doi.org/10.18653/v1/E17-1006

Hauk, O. (2016). Chapter 62 - What Does It Mean? A Review of the Neuroscientific Evidence for Embodied Lexical Semantics. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 777–788). Academic Press. https://doi.org/10.1016/B978-0-12-407794-2.00062-6

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic Representation of Action Words in Human Motor and Premotor Cortex. *Neuron*, *41*(2), 301–307. https://doi.org/10.1016/S0896-6273(03)00838-9

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Hess, D. J., Foss, D. J., & Carroll, P. (1995). Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology: General*, *124*(1), 62–82. https://doi.org/10.1037/0096-3445.124.1.62

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. Neural Computation. https://doi.org/10.1162/neco.1997.9.8.1735

Hoenig, K., Sim, E.-J., Bochev, V., Herrnberger, B., & Kiefer, M. (2008). Conceptual Flexibility in the Human Brain: Dynamic Recruitment of Semantic Maps from Visual, Motor, and Motion-related Areas. *Journal of Cognitive Neuroscience*, *20*(10), 1799–1814. https://doi.org/10.1162/jocn.2008.20123

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, *125*(3), 293–328. https://doi.org/10.1037/rev0000094

Hollenstein, N., de la Torre, A., Langer, N., & Zhang, C. (2019). CogniVal: A Framework for Cognitive Word Embedding Evaluation. *Proceedings of the 23rd Conference*

on Computational Natural Language Learning (CoNLL), 538–549. https://doi.org/10.18653/v1/K19-1050

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

Huettig, F., Olivers, C. N. L., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica*, *137*(2), 138–150. https://doi.org/10.1016/j.actpsy.2010.07.013

Humphreys, G. W., Price, C. J., & Riddoch, M. J. (1999). From objects to names: A cognitive neuroscience approach. *Psychological Research*, *62*(2–3), 118–130. https://doi.org/10.1007/s004260050046

Hutchinson, S., & Louwerse, M. M. (2013). What's Up can be Explained by Language Statistics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*. https://escholarship.org/uc/item/4h97698p

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171. https://doi.org/10.1016/j.jml.2015.10.007

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, *32*(8), 954–965. https://doi.org/10.1080/23273798.2016.1242761

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, *98*, 1–11. https://doi.org/10.1016/j.jml.2017.09.002

Jackendoff, R. (2003). Précis of Foundations of Language: Brain, Meaning, Grammar, Evolution,. *Behavioral and Brain Sciences*, *26*(6), 651–665. https://doi.org/10.1017/S0140525X03000153

Jamieson, R. K., Johns, B. T., Vokey, J. R., & Jones, M. N. (2022). Instance theory as a domain-general framework for cognitive psychology. *Nature Reviews Psychology*, *1*(3), 174–183. https://doi.org/10.1038/s44159-022-00025-3

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356

Jegerski, J. (2013). Self-Paced Reading. In *Research Methods in Second Language Psycholinguistics*. Routledge.

Johnson, B. P., Dayan, E., Censor, N., & Cohen, L. G. (2022). Crowdsourcing in Cognitive and Systems Neuroscience. *The Neuroscientist*, *28*(5), 425–437. https://doi.org/10.1177/10738584211017018

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552. https://doi.org/10.1016/j.jml.2006.07.003

Jordan, M. I. (1997). Chapter 25 - Serial Order: A Parallel Distributed Processing Approach. In J. W. Donahoe & V. Packard Dorsel (Eds.), *Advances in Psychology* (Vol. 121, pp. 471–495). North-Holland. https://doi.org/10.1016/S0166-4115(97)80111-2

Just, M. A., & Carpenter, P. A. (1980). A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review*, *87*(4), 329.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and Processes in Reading Comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228.

Kan, I. P., Barsalou, L. W., Olseth Solomon, K., Minor, J. K., & Thompson-Schill, S. L. (2003). Role of Mental Imagery in a Property Verification Task: fMRI Evidence for Perceptual Representations of Conceptual Knowledge. *Cognitive Neuropsychology*, *20*(3–6), 525–540. https://doi.org/10.1080/02643290244000257

Kaschak, M. P., & Glenberg, A. M. (2000). Constructing Meaning: The Role of Affordances and Grammatical Constructions in Sentence Comprehension. *Journal of Memory and Language*, *43*(3), 508–529. https://doi.org/10.1006/jmla.2000.2705

Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R. H., Aveyard, M., Blanchard, A. A., & Zwaan, R. A. (2005). Perception of motion affects language processing. *Cognition*, *94*(3), B79–B89. https://doi.org/10.1016/j.cognition.2004.06.005

Kemmerer, D., Miller, L., MacPherson, M. K., Huber, J., & Tranel, D. (2013). An investigation of semantic similarity judgments about action and non-action verbs in Parkinson's disease: Implications for the Embodied Cognition Framework. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00146

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee corpus. *Proceedings of the 12th European Conference on Eye Movement*.

Kennedy, A., Pynte, J., Murray, W. S., & Paul, S.-A. (2013). Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. *Quarterly Journal of*

Experimental Psychology, *66*(3), 601–618. https://doi.org/10.1080/17470218.2012.676054

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45. https://doi.org/10.1007/s10988-006-9008-0

Kennedy, C. (2012). Adjectives. In G. Russell & D. Graff Fara (Eds.), *Routledge companion to philosophy of language* (1st ed., pp. 328–341). Routledge. https://doi.org/10.4324/9780203206966

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805–825. https://doi.org/10.1016/j.cortex.2011.04.006

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

Konkle, T., & Oliva, A. (2012). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, *74*(6), 1114–1124. https://doi.org/10.1016/j.neuron.2012.04.036

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Kuhnke, P., Kiefer, M., & Hartwigsen, G. (2020). Task-Dependent Recruitment of Modality-Specific and Multimodal Regions during Conceptual Processing. *Cerebral Cortex*, *30*(7), 3938–3959. https://doi.org/10.1093/cercor/bhaa010

Kuperberg, G. R. (2021). Tea With Milk? A Hierarchical Generative Framework of Sequential Event Comprehension. *Topics in Cognitive Science*, *13*(1), 256–298. https://doi.org/10.1111/tops.12518

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A Look around at What Lies Ahead: Prediction and Predictability in Language Processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195395518.003.0065

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, *11*(2), 99–116. https://doi.org/10.1016/0301-0511(80)90046-0

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lake, B. M., & Murphy, G. L. (2023). Word meaning in minds and machines. *Psychological Review*, *130*(2), 401–431. https://doi.org/10.1037/rev0000297

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive Science*, *4*(2), 195–208.

Lakoff, G., & Johnson, M. (2008). *Metaphors We Live By*. University of Chicago Press.

Lakoff, G., Johnson, M., & Sowa, J. F. (1999). Review of Philosophy in the Flesh: The embodied mind and its challenge to Western thought. *Computational Linguistics*, *25*(4), 631–634.

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: Evidence from semantic dementia, HSVE and a neural network model. *Brain*, *130*(4), 1127–1137. Scopus. https://doi.org/10.1093/brain/awm025

Lambon Ralph, M. A., Mcclelland, J. L., Patterson, K., Galton, C. J., & Hodges, J. R. (2001). No right to speak? The relationship between object naming and semantic impairment: Neuropsychological evidence and a computational model. *Journal of Cognitive Neuroscience*, *13*(3), 341–356. Scopus. https://doi.org/10.1162/08989290151137395

Lambon Ralph, M. A., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences*, *107*(6), 2717–2722. https://doi.org/10.1073/pnas.0907307107

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Ledoux, K., Camblin, C. C., Swaab, T. Y., & Gordon, P. C. (2006). Reading Words in Discourse: The Modulation of Lexical Priming Effects by Message-Level Context. *Behavioral and Cognitive Neuroscience Reviews*, *5*(3), 107–127. https://doi.org/10.1177/1534582306289573

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1–31. https://www.italian-journal-linguistics.com/2008-2/

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*(Volume 4, 2018), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A., & Miliani, M. (2022). A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*, *56*(4), 1269–1313. https://doi.org/10.1007/s10579-021-09575-z

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38. https://doi.org/10.1017/S0140525X99001776

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In *Sentence processing* (pp. 78–114). Psychology Press.

Liashchynskyi, P., & Liashchynskyi, P. (2019). *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS* (arXiv:1912.06059). arXiv. https://doi.org/10.48550/arXiv.1912.06059

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Lopopolo, A., & Rabovsky, M. (2024). Tracking Lexical and Semantic Prediction Error Underlying the N400 Using Artificial Neural Network Models of Sentence Processing. *Neurobiology of Language*, *5*(1), 136–166. https://doi.org/10.1162/nol_a_00134

Lorch, R. F., & Myers, J. L. (1990). Regression Analyses of Repeated Measures Data in Cognitive Research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 149–157. https://doi.org/10.1037/0278-7393.16.1.149

Louwerse, M. M. (2007). Symbolic or embodied representations: A case for symbol interdependency. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 107–120). Psychology Press.

Louwerse, M. M. (2011). Symbol Interdependency in Symbolic and Embodied Cognition. *Topics in Cognitive Science*, *3*(2), 273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x

Louwerse, M. M. (2018). Knowing the Meaning of a Word by the Linguistic and Perceptual Company It Keeps. *Topics in Cognitive Science*, *10*(3), 573–589. https://doi.org/10.1111/tops.12349

Louwerse, M. M., & Connell, L. (2011). A Taste of Words: Linguistic Context and Perceptual Simulation Predict the Modality of Words. *Cognitive Science*, *35*(2), 381–398. https://doi.org/10.1111/j.1551-6709.2010.01157.x

Louwerse, M. M., & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, *114*(1), 96–104. https://doi.org/10.1016/j.cognition.2009.09.002

Louwerse, M. M., & Zwaan, R. A. (2009). Language Encodes Geographical Information. *Cognitive Science*, *33*(1), 51–73. https://doi.org/10.1111/j.1551-6709.2008.01003.x

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction. *Cognitive Science*, *42*, 1166–1183. https://doi.org/10.1111/cogs.12597

Lowe, W., & McDonald, S. A. (2000). The Direct Route: Mediated Priming in Semantic Space. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *22*. https://escholarship.org/content/qt9tf6q71r/qt9tf6q71r.pdf

Luck, S. J. (2012). Event-related potentials. In *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 523–546). American Psychological Association. https://doi.org/10.1037/13619-028

Lucy, L., & Gauthier, J. (2017). Are Distributional Representations Ready for the Real World? Evaluating Word Vectors for Grounded Perceptual Meaning. *Proceedings of the First Workshop on Language Grounding for Robotics*, 76–85. https://doi.org/10.18653/v1/W17-2810

Lund, K., & Burgess, C. (1996). Hyperspace analogue to language (HAL): A general model semantic representation. *Brain and Cognition*, *30*(3).

Luong, M.-T., Pham, H., & Manning, C. D. (2015). *Effective Approaches to Attention-based Neural Machine Translation* (arXiv:1508.04025). arXiv. https://doi.org/10.48550/arXiv.1508.04025

Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

Mahon, B. Z., & Caramazza, A. (2005). The orchestration of the sensory-motor systems: Clues from Neuropsychology. *Cognitive Neuropsychology*, *22*(3–4), 480–494. https://doi.org/10.1080/02643290442000446

Mahon, B. Z., & Caramazza, A. (2009). Concepts and Categories: A Cognitive Neuropsychological Perspective. *Annual Review of Psychology*, *60*(Volume 60, 2009), 27–51. https://doi.org/10.1146/annurev.psych.60.110707.163532

Marcot, B. G., & Hanea, A. M. (2021). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics*, *36*(3), 2009–2031. https://doi.org/10.1007/s00180-020-00999-9

Marslen-Wilson, W. D. (1973). Linguistic Structure and Speech Shadowing at Very Short Latencies. *Nature*, *244*(5417), 522–523. https://doi.org/10.1038/244522a0

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Martin, A. (2007). The Representation of Object Concepts in the Brain. *Annual Review of Psychology*, *58*(Volume 58, 2007), 25–45. https://doi.org/10.1146/annurev.psych.57.102904.190143

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology*, *11*(2), 194–201. https://doi.org/10.1016/S0959-4388(00)00196-3

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

McDonald, S. A., & Shillcock, R. C. (2003). Eye Movements Reveal the On-Line Computation of Lexical Probabilities During Reading. *Psychological Science*, *14*(6), 648–652. https://doi.org/10.1046/j.0956-7976.2003.psci_1480.x

McNamara, T., P., & Holbrook, J. B. (2003). Semantic memory and priming. In *Handbook of Psychology, Experimental Psychology* (Vol. 4, pp. 447–474). John Wiley & Sons.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. https://doi.org/10.3758/BF03192726

Medler, D. A., & Binder, J. R. (2005). *MCWord: An On-Line Orthographic Database of the English Language*. http://www.neuro.mcw.edu/mcword/

Merkx, D., & Frank, S. L. (2021). Human Sentence Processing: Recurrence or Attention? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 12–22. https://doi.org/10.18653/v1/2021.cmcl-1.2

Merrill, W., Goldberg, Y., Schwartz, R., & Smith, N. A. (2021). Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *Transactions of the Association for Computational Linguistics*, *9*, 1047–1060. https://doi.org/10.1162/tacl_a_00412

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*(7), 788–804. https://doi.org/10.1016/j.cortex.2010.11.002

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

Michaelov, J. A., Arnett, C., & Bergen, B. K. (2024). *Revenge of the Fallen? Recurrent Models Match Transformers at Predicting Human Language Comprehension Metrics* (arXiv:2404.19178). arXiv. https://doi.org/10.48550/arXiv.2404.19178

Michaelov, J. A., Bardolph, M. D., Coulson, S., & Bergen, B. (2021). Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). https://escholarship.org/uc/item/9z06m20f

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, *5*(1), 107–135. https://doi.org/10.1162/nol_a_00105

Michaelov, J. A., & Bergen, B. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? *Proceedings of the 24th Conference on Computational Natural Language Learning*, 652–663. https://doi.org/10.18653/v1/2020.conll-1.53

Michaelov, J. A., & Bergen, B. (2022). *The more human-like the language model, the more surprisal is the best predictor of N400 amplitude*. NeurIPS 2022 Workshop

on Information-Theoretic Principles in Cognitive Systems. https://openreview.net/forum?id=uCgYvb8GNQZ

Michaelov, J. A., Coulson, S., & Bergen, B. (2023a). Can Peanuts Fall in Love with Distributional Semantics? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45). https://escholarship.org/uc/item/08h921zh

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2023b). So Cloze Yet So Far: N400 Amplitude Is Better Predicted by Distributional Information Than Human Predictability Judgements. *IEEE Transactions on Cognitive and Developmental Systems*, *15*, 1033–1042. https://doi.org/10.1109/TCDS.2022.3176783

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

Misra, K., Ettinger, A., & Rayz, J. (2020). Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4625–4635. https://doi.org/10.18653/v1/2020.findings-emnlp.415

Mitchell, J., & Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, *34*(8), 1388–1429. https://doi.org/10.1111/j.1551-6709.2010.01106.x

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 398–408. https://aclanthology.org/E12-1041.pdf

Morris, R. K. (2006). Lexical processing and sentence context effects. In *Handbook of psycholinguistics* (pp. 377–401). Academic Press.

Munikar, M., Shakya, S., & Shrestha, A. (2019). Fine-grained Sentiment Classification using BERT. *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, 1–5. https://doi.org/10.1109/AITB48515.2019.8947435

Murphy, G. L. (1988). Comprehending Complex Concepts. *Cognitive Science*, *12*(4), 529–562. https://doi.org/10.1207/s15516709cog1204_2

Murphy, G. L. (2004). *The Big Book of Concepts*. MIT Press.

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*(2–3), 131–157. https://doi.org/10.1080/01638539809545042

Myers, J. L., O'Brien, E. J., Albrecht, J. E., & Mason, R. A. (1994). Maintaining Global Coherence During Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 876–886. https://doi.org/10.1037/0278-7393.20.4.876

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Matthew Husband, E., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., … Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a

large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

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

Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science*, *13*(6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05

O'Brien, E. J., & Albrecht, J. E. (1992). Comprehension Strategies in the Development of a Mental Model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 8.

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of Structural Parsers and Neural Language Models as Surprisal Estimators. *Frontiers in Artificial Intelligence*, *5*, 777963. https://doi.org/10.3389/frai.2022.777963

Oh, B.-D., & Schuler, W. (2023). Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of*

the Association for Computational Linguistics, *11*, 336–350. https://doi.org/10.1162/tacl_a_00548

Olson, D. R., & Filby, N. (1972). On the comprehension of active and passive sentences. *Cognitive Psychology*, *3*(3), 361–381. https://doi.org/10.1016/0010-0285(72)90013-8

Ontanon, S., Ainslie, J., Fisher, Z., & Cvicek, V. (2022). Making Transformers Solve Compositional Tasks. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3591–3607. https://doi.org/10.18653/v1/2022.acl-long.251

Otten, M., Nieuwland, M. S., & Van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, *8*(1), 89. https://doi.org/10.1186/1471-2202-8-89

Otten, M., & Van Berkum, J. J. A. (2008). Discourse-Based Word Anticipation During Language Processing: Prediction or Priming? *Discourse Processes*, *45*(6), 464–496. https://doi.org/10.1080/01638530802356463

Paivio, A. (1990). *Mental Representations: A Dual Coding Approach*. Oxford University Press.

Partee, B. H. (2007). Compositionality and coercion in semantics: The dynamics of adjective meaning. *Cognitive Foundations of Interpretation*, 145–161. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=98e6a1050131c270703f9bd0e17678ae1ab9d40a#page=153

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019).

PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Patterson, K., & Lambon Ralph, M. A. (2016). Chapter 61—The Hub-and-Spoke Hypothesis of Semantic Memory. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 765–775). Academic Press. https://doi.org/10.1016/B978-0-12-407794-2.00061-4

Peelle, J. E., Miller, R. L., Rogers, C. S., Spehar, B., Sommers, M. S., & Van Engen, K. J. (2020). Completion norms for 3085 English sentence contexts. *Behavior Research Methods*, *52*(4), 1795–1799. https://doi.org/10.3758/s13428-020-01351-1

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3–4), 175–190. https://doi.org/10.1080/02643294.2016.1176907

Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processes*, *25*(2–3), 363–377. https://doi.org/10.1080/01638539809545033

Piantadosi, S., & Hill, F. (2022, October 21). *Meaning without reference in large language models*. NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI). https://openreview.net/forum?id=nRkJEwmZnM

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191–199.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*(1), 77–105. https://doi.org/10.1016/0004-3702(90)90005-K

Posit Team. (2024). *RStudio: Integrated Development Environment for R* [Computer software]. Posit Software, PBC. http://www.posit.co/

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, *22*(2), 253–279. https://doi.org/10.1017/S0140525X9900182X

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(7), 576–582. https://doi.org/10.1038/nrn1706

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, *3*(1), 111–132. https://doi.org/10.1017/S0140525X00002053

Quante, L., Bölte, J., & Zwitserlood, P. (2018). Dissociating predictability, plausibility and possibility of sentence continuations in reading: Evidence from late-positivity ERPs. *PeerJ*, *6*, e5717. https://doi.org/10.7717/peerj.5717

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, *12*(5), 410–430. https://doi.org/10.1002/bs.3830120511

Quillian, M. R. (1969). The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM*, *12*(8), 459–476. https://doi.org/10.1145/363196.363214

R Core Team. (2018). *R Foundation for Statistical Computing; Vienna, Austria: 2014*. [Computer software]. https://www.r-project.org/

Rabovsky, M., & McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: A neural network-based approach to capturing neural responses during human language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190313. https://doi.org/10.1098/rstb.2019.0313

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, *1*(8), 9.

Riordan, B., & Jones, M. N. (2011). Redundancy in Perceptual and Linguistic Experience: Comparing Feature-Based and Distributional Models of Semantic Representation. *Topics in Cognitive Science*, *3*(2), 303–345. https://doi.org/10.1111/j.1756-8765.2010.01111.x

Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., & Adler, A. (2016). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language*, *87*, 16–37. https://doi.org/10.1016/j.jml.2015.10.006

Rodd, J. M., Lopez Cutrin, B., Kirsch, H., Millar, A., & Davis, M. H. (2013). Long-term priming of the meanings of ambiguous words. *Journal of Memory and Language*, *68*(2), 180–198. https://doi.org/10.1016/j.jml.2012.08.002

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press.

Rogers, T. T., & Wolmetz, M. (2016). Conceptual knowledge representation: A cross-section of current research. *Cognitive Neuropsychology*, *33*(3–4), 121–129. https://doi.org/10.1080/02643294.2016.1188066

Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia*, *51*(3), 437–447. https://doi.org/10.1016/j.neuropsychologia.2012.12.002

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rumelhart, D. E. (1980). Schemata: The Building Blocks of Cognition. In *Theoretical Issues in Reading Comprehension* (1st ed.). Routledge. https://doi.org/10.4324/9781315107493

Rumelhart, D. E., & McClelland, J. L. (1987). A General Framework for Parallel Distributed Processing. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (pp. 45–76). Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations. MIT Press. https://ieeexplore.ieee.org/abstract/document/6302935

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces* [Doctoral dissertation]. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A189276&dswid=-3486

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*, 33–53.

Santos, A., Chaigneau, S. E., Simmons, W. K., & Barsalou, L. W. (2011). Property generation reflects word association and situated simulation. *Language and Cognition*, *3*(1), 83–119. https://doi.org/10.1515/langcog.2011.004

Schustack, W., Ehrlich, S. E., & Rayner, K. (1987). Local and Global Sources of Contextual Facilitation. *Journal of Memory and Language*, *26*(3), 322–340. https://doi.org/10.1016/0749-596X(87)90117-3

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 344–354. https://doi.org/10.1037/0278-7393.14.2.344

Schwanenflugel, P. J., & White, C. R. (1991). The Influence of Paragraph Information on the Processing of Upcoming Words. *Reading Research Quarterly*, *26*(2), 160–177. https://doi.org/10.2307/747980

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

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147. https://doi.org/10.1016/S0010-0277(99)00025-6

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121. https://doi.org/10.1073/pnas.2307876121

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shapiro, L., & Spaulding, S. (2024). Embodied Cognition. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2024/entries/embodied-cognition/

Shwartz, V., & Dagan, I. (2019). Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition. *Transactions of the Association for Computational Linguistics*, *7*, 403–419. https://doi.org/10.1162/tacl_a_00277

Simmons, W. K., Ramjee, V., Beauchamp, M. S., McRae, K., Martin, A., & Barsalou, L. W. (2007). A common neural substrate for perceiving and knowing about color. *Neuropsychologia*, *45*(12), 2802–2810. https://doi.org/10.1016/j.neuropsychologia.2007.05.002

Smith, E. E., & Osherson, D. N. (1984). Conceptual Combination with Prototype Concepts. *Cognitive Science*, *8*(4), 337–361. https://doi.org/10.1207/s15516709cog0804_2

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining Prototypes: A Selective Modification Model. *Cognitive Science*, *12*(4), 485–527. https://doi.org/10.1207/s15516709cog1204_1

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3), 214–241. https://doi.org/10.1037/h0036351

Smith, L. B., Cooney, N. J., & McCord, C. (1986). *What Is 'High'? The Development of Reference Points for 'High' and 'Low'*. 21.

Smith, N. J., & Levy, R. (2008). Optimal Processing Times in Reading: A Formal Model and Empirical Investigation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*. https://escholarship.org/uc/item/3mr8m3rf

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1), 159–216. https://doi.org/10.1016/0004-3702(90)90007-M

Solomon, S. H., & Thompson-Schill, S. L. (2020). Feature Uncertainty Predicts Behavioral and Neural Responses to Combined Concepts. *Journal of Neuroscience*, *40*(25), 4900–4912. https://doi.org/10.1523/JNEUROSCI.2926-19.2020

Solt, S. (2019). Adjective Meaning and Scales. In C. Cummins & N. Katsos (Eds.), *The Oxford Handbook of Experimental Semantics and Pragmatics* (pp. 262–282). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198791768.013.27

Sommerauer, P., & Fokkens, A. (2018). Firearms and Tigers are Dangerous, Kitchen Knives and Zebras are Not: Testing whether Word Embeddings Can Tell.

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 276–286. https://doi.org/10.18653/v1/W18-5430

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1), Article 1. https://doi.org/10.1609/aaai.v31i1.11164

Springer, K., & Murphy, G. L. (1992). Feature Availability in Conceptual Combination. *Psychological Science*, *3*(2), 111–117. https://doi.org/10.1111/j.1467-9280.1992.tb00008.x

Stanfield, R. A., & Zwaan, R. A. (2001). The Effect of Implied Orientation Derived from Verbal Context on Picture Recognition. *Psychological Science*, *12*(2), 153–156. https://doi.org/10.1111/1467-9280.00326

Stanovich, K. E., & West, R. F. (1981). The Effect of Sentence Context on Ongoing Word Recognition: Tests of a Two-Process Theory. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(3), 658–672. https://doi.org/10.1037/0096-1523.7.3.658

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, *112*(1), 1–36. https://doi.org/10.1037/0096-3445.112.1.1

Staub, A. (2015). The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation: Predictability and Eye Movements. *Language and Linguistics Compass*, *9*(8), 311–327. https://doi.org/10.1111/lnc3.12151

Stewart, A. J., Kidd, E., & Haigh, M. (2009). Early Sensitivity to Discourse-Level Anomalies: Evidence From Self-Paced Reading. *Discourse Processes*, *46*(1), 46–69. https://doi.org/10.1080/01638530802629091

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645–659. https://doi.org/10.1016/S0022-5371(79)90355-4

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, *123*, 104311. https://doi.org/10.1016/j.jml.2021.104311

Tabossi, P. (1988). Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory and Language*, *27*(3), 324–340. https://doi.org/10.1016/0749-596X(88)90058-7

Tabossi, P., & Johnson-Laird, P. N. (1980). Linguistic Context and the Priming of Semantic Information. *Quarterly Journal of Experimental Psychology*, *32*(4), 595–603. https://doi.org/10.1080/14640748008401848

Tanaka, K. (1996). Inferotemporal Cortex and Object Vision. *Annual Review of Neuroscience*, *19*(1), 109–139. https://doi.org/10.1146/annurev.ne.19.030196.000545

Taylor, W. L. (1953). “Cloze Procedure”: A New Tool for Measuring Readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International*

Journal of Psychophysiology, *83*(3), 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007

Traxler, M. J., & Foss, D. J. (2000). Effects of sentence constraint on priming in natural language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1266–1282. https://doi.org/10.1037/0278-7393.26.5.1266

Traxler, M. J., Foss, D. J., Seely, R. E., Kaup, B., & Morris, R. K. (2000). Priming in Sentence Processing: Intralexical Spreading Activation, Schemas, and Situation Models. *Journal of Psycholinguistic Research*, *29*(6), 581–595.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*(3), 285–318. https://doi.org/10.1006/jmla.1994.1014

Turton, J., Smith, R. E., & Vinson, D. (2021). Deriving Contextualised Semantic Features from BERT (and Other Transformer Model) Embeddings. *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, 248–262. https://doi.org/10.18653/v1/2021.repl4nlp-1.26

Turton, J., Vinson, D., & Smith, R. (2020). Extrapolating Binder Style Word Embeddings to New Words. *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, 1–8. https://aclanthology.org/2020.lincr-1.1/

Utsumi, A. (2020). Exploring What Is Encoded in Distributional Word Vectors: A Neurobiologically Motivated Analysis. *Cognitive Science*, *44*(6). https://doi.org/10.1111/cogs.12844

Van Berkum, J. J. A., 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 Dam, W. O., van Dijk, M., Bekkering, H., & Rueschemeyer, S.-A. (2012). Flexibility in embodied lexical-semantic representations. *Human Brain Mapping*, *33*(10), 2322–2333. https://doi.org/10.1002/hbm.21365

Van Dam, W., Rueschemeyer, S.-A., Lindemann, O., & Bekkering, H. (2010). Context Effects in Embodied Lexical-Semantic Processing. *Frontiers in Psychology*, *1*. https://www.frontiersin.org/articles/10.3389/fpsyg.2010.00150

Van Dantzig, S., Cowell, R. A., Zeelenberg, R., & Pecher, D. (2011). A sharp image or a sharp knife: Norms for the modality-exclusivity of 774 concept-property items. *Behavior Research Methods*, *43*(1), 145–154. https://doi.org/10.3758/s13428-010-0038-8

Van Dantzig, S., Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2008). Perceptual Processing Affects Conceptual Processing. *Cognitive Science*, *32*(3), 579–590. https://doi.org/10.1080/03640210802035365

van Dijk, T. A. (1999). Context Models in Discourse Processing. In *The construction of mental representations during reading* (pp. 123–148).

van Dijk, T. A. (2006). Discourse, context and cognition. *Discourse Studies*, *8*(1), 159–177. https://doi.org/10.1177/1461445606059565

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly*

Journal of Experimental Psychology, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

van Schijndel, M., & Linzen, T. (2018). *Can Entropy Explain Successor Surprisal Effects in Reading?* (arXiv:1810.11481). arXiv. https://doi.org/10.48550/arXiv.1810.11481

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vecchi, E. M., Marelli, M., Zamparelli, R., & Baroni, M. (2017). Spicy Adjectives and Nominal Donkeys: Capturing Semantic Deviance Using Compositionality in Distributional Spaces. *Cognitive Science*, *41*(1), 102–136. https://doi.org/10.1111/cogs.12330

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, *56*(3), 229–255. https://doi.org/10.1080/0163853X.2018.1448677

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space

hypothesis. *Cognitive Psychology*, *48*(4), 422–488. https://doi.org/10.1016/j.cogpsych.2003.09.001

Visser, M., Jefferies, E., & Lambon Ralph, M. A. (2010). Semantic Processing in the Anterior Temporal Lobes: A Meta-analysis of the Functional Neuroimaging Literature. *Journal of Cognitive Neuroscience*, *22*(6), 1083–1094. https://doi.org/10.1162/jocn.2009.21309

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., & Korhonen, A. (2020). Probing Pretrained Language Models for Lexical Semantics. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7222–7240). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.586

Wang, X., & Bi, Y. (2021). Idiosyncratic Tower of Babel: Individual Differences in Word-Meaning Representation Increase as Word Abstractness Increases. *Psychological Science*, *32*(10), 1617–1635. https://doi.org/10.1177/09567976211003877

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, & R. Cotterell (Eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (pp. 1–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-babylm.1

Wicha, N. Y. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, *346*(3), 165–168. https://doi.org/10.1016/S0304-3940(03)00599-8

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating Words and Their Gender: An Event-related Brain Potential Study of Semantic Integration, Gender Expectancy, and Gender Agreement in Spanish Sentence Reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. https://doi.org/10.1162/0898929041920487

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). *On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior* (arXiv:2006.01912). arXiv. http://arxiv.org/abs/2006.01912

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470. https://doi.org/10.1162/tacl_a_00612

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Willems, R. M., & Peelen, M. V. (2021). How context changes the neural basis of perception and language—ScienceDirect. *Iscience*, *24*(5), 102392. https://www.sciencedirect.com/science/article/pii/S2589004221003606

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

Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K., & Barsalou, L. W. (2011). Grounding emotion in situated conceptualization. *Neuropsychologia*, *49*(5), 1105–1127. https://doi.org/10.1016/j.neuropsychologia.2010.12.032

Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin & Review*, *4*(2), 167–183. https://doi.org/10.3758/BF03209392

Wisniewski, E. J., & Love, B. C. (1998). Relations versus Properties in Conceptual Combination. *Journal of Memory and Language*, *38*(2), 177–202. https://doi.org/10.1006/jmla.1997.2550

Wittgenstein, L. (1953). *Philosophical Investigations*. John Wiley & Sons.

Wlotko, E. W., & Federmeier, K. D. (2007). Finding the right word: Hemispheric asymmetries in the use of sentence context information. *Neuropsychologia*, *45*(13), 3001–3014. https://doi.org/10.1016/j.neuropsychologia.2007.05.013

Wlotko, E. W., & Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage*, *62*(1), 356–366. https://doi.org/10.1016/j.neuroimage.2012.04.054

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). *HuggingFace's*

Transformers: State-of-the-art Natural Language Processing (arXiv:1910.03771). arXiv. http://arxiv.org/abs/1910.03771

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

Yee, E., Ahmed, S. Z., & Thompson-Schill, S. L. (2012). Colorless Green Ideas (Can) Prime Furiously. *Psychological Science*, *23*(4), 364–369. https://doi.org/10.1177/0956797611430691

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, *23*(4), 1015–1027. https://doi.org/10.3758/s13423-015-0948-7

Zarcone, A., Van Schijndel, M., Vogels, J., & Demberg, V. (2016). Salience and Attention in Surprisal-Based Accounts of Language Processing. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00844

Zeki, S., Watson, J. D., Lueck, C. J., Friston, K. J., Kennard, C., & Frackowiak, R. S. (1991). A direct demonstration of functional specialization in human visual cortex. *Journal of Neuroscience*, *11*(3), 641–649. https://doi.org/10.1523/JNEUROSCI.11-03-00641.1991

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, *23*(4), 1028–1034. https://doi.org/10.3758/s13423-015-0864-x

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*(5), 292–297. https://doi.org/10.1111/j.1467-9280.1995.tb00513.x

Zwaan, R. A., & Radvansky, G. A. (1998). Situation Models in Language Comprehension and Memory. *Psychological Bulletin*, *123*, 24.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language Comprehenders Mentally Represent the Shapes of Objects. *Psychological Science*, *13*(2), 168–171. https://doi.org/10.1111/1467-9280.00430