SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Psychology

PhD program in Psychology, Linguistics and Cognitive Neuroscience Cycle XXXVII

Curriculum in Mind, Brain and Behavior

MULTIPLE ROUTES TO CONCEPTUAL COMBINATION: A DISTRIBUTIONAL MODELING APPROACH

Surname: Ciapparelli Name: Marco

Registration number: 815165

Tutor: Prof. Marco Marelli

Co-tutor: Prof. Carlo Reverberi

Coordinator: Prof.ssa Simona Sacchi

ACADEMIC YEAR 2023/2024

Abstract

A core feature of human cognition is *conceptual combination,* namely the ability to build complex meanings (e.g., “swordfish”) from simpler constituents (“sword”, “fish”) whose relationship is implicit and must be inferred (“a fish with a jaw shaped like a sword”). Different theories claim that conceptual combination relies on multiple routes to meaning, resulting in alternative representations of the same concept. Specifically, concepts may be represented as both independent and morphologically bound constituents (e.g., “fish” and the head constituent “–fish”), depend on both language and sensorimotor information (e.g., the way “sword” and “fish” are used as words as well as their distinct perceptual and motor features), and result from long-term memory retrieval or active composition (e.g., “swordfish” as a single lexicalized unit or as the combination of “sword” and “fish”). However, many of these constructs remain elusive to operationalize and are, therefore, difficult to test experimentally. The present dissertation proposes to address this challenge by means of computational models of conceptual combination based on distributional semantics.

Chapter 2 leverages the intermediate representations of a compositional model to operationalize morphologically bound units (e.g., “sword” as the modifier “sword–”) and disentangle inhibitory and facilitatory dimensions of compound meaning. These are shown to predict eye movements in interaction with measures of reading experience, supporting the claim that morphologically bound units are developed with prolonged exposure to compounding and affect how compounds are processed. Leveraging the similarity between language and vision, Chapter 3 tests the hypothesis that the visual features of novel combinations (e.g., “cactus fish”) can be predicted by mapping the representations of a (compositional) language model to those of a vision model. Synthetic images depicting novel combinations are generated, and participants are tasked with selecting those that best match their intuitions. Language-to-vision mapping mostly succeeds in predicting participants’ behavior, but the superior performance of the head constituent (over a compositional model) suggests that language provides only a simple pointer to the perceptual features of novel combinations, calling for additional combinatorial processes. Chapter 4 consists of a re-analysis of four fMRI studies employing two-word combinations but differing in task requirements. “Lexicalized” and “compositional” representations are compared in terms of their degree of match with neural activity from target regions of interest. Evidence for lexicalized representations is absent, while converging evidence from confirmatory and exploratory analyses points to a cluster of compositional representations in the left inferior frontal gyrus, showing that this region represents combinatorial information beyond task demands. Finally, chapter 5 shows that models from two

classes – simpler compositional models and large language models – can reveal the implicit semantic relations that link constituent words (e.g., “snowball” is “a ball *made of* snow”). Large language models are markedly superior in the case of familiar compounds. However, their performance on novel compounds experiences a substantial drop, falling below that of simpler compositional models, suggesting that large language models struggle with stimuli that require genuine conceptual combination.

Together, these contributions show that, by defining both the representations of individual concepts and the functions that combine them, compositional models based on distributional semantics provide a window into representations and processes that cannot be directly examined, characterizing alternative routes to complex meaning which are based on different representations of the same constituent, different representations of the same combination, and different sources of information.

Contents

1. **Chapter 1: Introduction 8**
   - **1.1. Concepts 10**
     - 1.1.1. Theories and models of semantic representations 10
     - 1.1.2. Distributional semantics models 13
     - 1.1.3. Large language models 16
     - 1.1.4. Data-driven models beyond language 17
   - **1.2. Conceptual combination 19**
     - 1.2.1. Theories of conceptual combination 19
     - 1.2.2. The neural bases of conceptual combination 21
     - 1.2.3. The role of morphological structure 22
     - 1.2.4. The role of sensorimotor information 23
     - 1.2.5. The interplay between storage and computation 24
   - **1.3. Computational models of conceptual combination 25**
     - 1.3.1. Motivating computational implementations of conceptual combination 25
     - 1.3.2. Compositional distributional semantics models 26
     - 1.3.3. Conceptual combination in large language models 30
   - **1.4. The current dissertation 32**
2. **Chapter 2: Semantic transparency and reading experience 34**
   - **2.1. Introduction 34**
   - **2.2. Methods 38**
     - 2.2.1. Materials 38
     - 2.2.2. Modeling 38
     - 2.2.3. Analyses 39
   - **2.3. Results 45**
   - **2.4. Discussion 49**
3. **Chapter 3: Predicting conceptual combination by mapping language embeddings to visual representations 53**
   - **3.1. Introduction 53**
   - **3.2. Methods 56**
     - 3.2.1. Models 56
     - 3.2.2. Synthetic images validation 57

3.2.3. Language-to-vision mapping …… 61
3.2.4. Experiment ……... 64
3.3. Results …… 70
3.3.1. Descriptive results …… 70
3.3.2. Analysis 1: Predicting interpretation type …… 72
3.3.3. Analysis 2: Predicting typicality …….. 74
3.3.4. Analysis 3: Predicting selection times …….. 74
3.3.5. Exploratory analyses …… 75
3.4. Discussion ……... 78
4. Chapter 4: The neural bases of semantic composition: A model-based fMRI re-analysis …… 82
4.1. Introduction ……... 82
4.2. Materials and methods …….. 85
4.2.1. Experimental materials …… 85
4.2.2. Imaging methods …….. 87
4.2.3. Theoretical models ……... 90
4.3. Analyses and results …….. 93
4.3.1. Core ROIs …….… 93
4.3.2. Searchlight RSA ……... 95
4.4. Discussion ……... 97
5. Chapter 5: Compounding in large language models ……. 101
5.1. Introduction ……. 101
5.2. Methods …… 103
5.2.1. Large language models ……... 103
5.2.2. Large language models: Modeling and predicting relational interpretations ……. 104
5.2.3. Baseline models …….. 107
5.2.4. Baseline models: Modeling and predicting relational interpretations …… 109
5.3. Models evaluation …… 109
5.3.1. Analysis 1: Existing compounds rated in isolation …… 109
5.3.2. Analysis 2: Existing compounds rated in context ……... 112
5.3.3. Analysis 3: Novel compounds rated in context …….. 114
5.4. Discussion …… 116
6. Chapter 6: General discussion …… 120

6.1. Summary of contributions …….. 120
6.2. How can we model conceptual combination? Comparing different solutions ……... 124
6.3. Future work ……. 127
7. Appendix ……...… 131
7.1. Appendix A.1 …… 131
7.2. Appendix A.2 …… 139
7.3. Appendix A.3 …… 141
8. References ……. 146

Chapter 1

Introduction

> *"Nothing, at first view, may seem more unbounded than the thought of man, which not only escapes all human power and authority, but is not even restrained within the limits of nature and reality. To form monsters, and join incongruous shapes and appearances, costs the imagination no more trouble than to conceive the most natural and familiar objects. […] But though our thought seems to possess this unbounded liberty, we shall find, upon a nearer examination, that it is really confined within very narrow limits, and that all this creative power of the mind amounts to no more than the faculty of compounding, transposing, augmenting, or diminishing the materials afforded us by the senses and experience."*
>
> (Hume, 1777/2007, *An Enquiry Concerning Human Understanding*, Section II, p. 13)

One of the core features of human cognition is *conceptual combination,* namely the ability to construct complex meanings from simpler constituents. Consider the very example that Hume provides to illustrate his passage, quoted above: "When we think of a golden mountain, we only join two consistent ideas, *gold*, and *mountain*, with which we were formerly acquainted." (Hume, 1777/2007, *An Enquiry Concerning Human Understanding*, Section II, p. 13). Because the adjective "golden", taken literally, indicates something "made of gold", a straightforward interpretation of *golden mountain* is "a mountain made of gold". However, upon closer reflection, "golden" does not always indicate that something is "made of gold". For example, a strict interpretation of "golden" would mislead us into thinking that a "golden eagle" is "an eagle made of gold" and a "golden retriever" is "a dog made of gold". Indeed, in combination with a noun, "golden" can often simply indicate the color, not the material, of gold, an observation that might lead us to retract our initial interpretation of *golden mountain* in favor of "a yellow mountain". Drawing from our experience with "golden–" as a modifier, we might notice a systematic usage pattern whereby "golden" means that something is of the greatest quality, to be admired; that is the case for adjective-noun phrases like "golden voice", "golden boy", and "golden age".

The problem of conceptual combination only worsens when looking at noun-noun compound words. Indeed, while often ambiguous, the adjective in adjective-noun phrases serves the purpose of

specifying a property of the noun. Instead, the modifier (in English, the left constituent) of noun-noun compounds can not only specify an attribute of the head, but engage with the meaning of the head via extrinsic relations. For example, in "gold digger", the modifier "gold" does not assign a property to the head "digger"; instead, a "gold digger" is "someone who digs or mines for gold" or, in the more frequent metaphorical sense, "a person that forms a relationship with someone in order to gain wealth". While some interpretations are more salient and plausible than others, this does not preclude us from considering alternatives that might become more salient and plausible provided an adequate context. For example, we can easily imagine that a "gold digger" might be "a tool used to extract gold"; if we were miners and a colleague lamented the poor quality of "gold diggers", that interpretation would indeed be more plausible than "a person that forms a relationship with someone in order to gain wealth". Moreover, conceptual combination is not a "closed operation", that is, properties that are not true (or typical) of either constituent can be true (or typical) of the compound (e.g., "losing money" is not a typical property of either "empty things" or "stores", while it's a typical property of "empty stores"; Murphy, 1988). Such "emergent properties" (Lake & Murphy, 2021) depend on background knowledge that "emerges" during constituents interaction, and that can be multimodal in nature. For example, when generating visual features for "rolled-up lawn", people generated attributes, such as "roots" and "dirt", that would typically be occluded in the visual characterization of "lawn" (Wu & Barsalou, 2009). As this example hints at, conceptual combination relies on sources of information other than language (Lynott & Connell, 2010).

For cognitive scientists, adjective-noun phrases such as "golden mountain" and, most notably, noun-noun compound words like "gold digger" provide a privileged window to the study of conceptual combination: Because the relationship that links constituent words is implicit (e.g., "a mountain *made of* gold"), speakers must infer it based on a relatively impoverished input, i.e., they have to combine constituent concepts. In this sense, theories of conceptual combination attempt to explain the mechanism underlying this inference, its brain bases, and its observable consequences on behavior. However, as the above examples suggest, concepts appear to be combined quite erratically, perhaps too erratically to be considered the outcome of a systematic mechanism. Despite this complexity, compounding is a productive word formation process in languages such as English and German (Libben, 2010). The fact that speakers of those languages routinely generate novel combinations implies that compounding must be reliable enough for hearers to infer all implicit information. As will be discussed in section 1.2 it has been claimed that compounding, in order to be effective, must leverage multiple sources of information to maximize the opportunities of successful comprehension:

(i) *Morphological information:* Constituents are often part of productive morphological families (e.g., the familiar "golden–" modifier, or the head "–berry"). It has been argued that the repeated exposure to words *as constituents* could lead to the development of morphologically-dependent representations (Libben, 2014). The as-constituent representation would then guide the interpretation of novel compounds, especially when their behavior is systematic – as in "golden voice", "golden boy", and "golden age" (section 1.2.3);

(ii) *Multimodal information:* Both sensorimotor and language information interact to determine the appropriate combinations (Lynott & Connell, 2010; section 1.2.4);

(iii) *Storage and (different types of) computation:* Multiple interpretations of the same compound are automatically attempted together with the retrieval of a lexicalized meaning stored in long-term memory (Libben, 2010; section 1.2.5).

A fourth, overarching issue concerns how the brain supports these forms of combination (section 1.2.2). The present work focuses on these issues by relying on deep-learning models of semantics – mostly based on distributional patterns of language usage, but also on visual information. More specifically, the goal of this dissertation is to demonstrate that computational models of conceptual combination and compounding, by formalizing both the representations of constituent concepts and the function(s) that combines them, are well suited to characterize (i) different representations of the same constituent, (ii) different representations of the same combination, and (iii) different sources of information on which these representations are based (language- or vision-based). This is important to address the issues described above, which posit, for the same concept, representations that are implicit, come from different sources (linguistic and multimodal), and are the outcome of different computations.

In the remaining parts of this general introduction, I will first overview theories of conceptual representation and the main modelling framework adopted, i.e., distributional semantics (section 1.1). Then, theories of conceptual combination and its brain bases are described, as well as the role played by morphological and sensorimotor information, and how retrieval and combination processes might operate in parallel (section 1.2). Finally, I describe computational models of meaning combination and introduce how these are used to tackle the above challenges in the subsequent chapters (section 1.3).

1.1 Concepts

1.1.1 Theories and models of semantic representations

Concepts are mental representations of categories that are critical in drawing inferences, communicating, and forming preferences (Markman & Rein, 2013). They: "tie our past experiences to our present interactions with the world […] telling us what things there are and what properties they have" (p.1, Murphy, 2004). Many theoretical approaches have been proposed to characterize the structure and format of conceptual representations. Historically, the *classical view* (described in Smith and Medin, 1981), conceived concepts as mental representations of category definitions, namely sets of characteristics that are both necessary and sufficient to determine membership to a category. According to the classical view, all concepts could be sharply distinguished as being either inside a category or outside of it, and, once in a category, different concepts simply "belonged" to that category. However, empirical evidence showed that category membership is not a dichotomous property, and that different members of the same category can be "better" or "worse" representatives of that category, i.e., have different degrees of *typicality* (e.g., "dog" is a typical member of the "mammals" category, while "dolphin" is an atypical one). Tasks that involve relating an item to a category were found to be robustly influenced by item typicality, e.g., speeded and unspeeded categorization, category learning, category-based induction, and rating tasks (Murphy, 2004). Because the classical theory was incapable of accounting for these phenomena, scholars proposed alternative theories. According to the *prototype view* (Rosch and Mervis, 1975), categories are centered on their typical representation, called "prototype", and new concepts become members of a category if they are sufficiently close or similar to its prototype. Instead, the *exemplar view* (Medin and Schaffer, 1978) rejects the idea that there are single representations encompassing entire concepts (e.g., prototypes); rather, concepts are sets of exemplars present in memory. For something to belong to a category, then, it needs to be more similar to a certain set of exemplars than others. Central to these theories is the notion of *semantic similarity*, a property called upon to account for a vast array of empirical observations such as typicality effects, semantic priming and neuropsychological findings (Vigliocco & Vinson, 2012). To account for semantic memory phenomena and theoretical predictions, one needs an account of how concepts are represented. In this regard, different accounts, or classes of models, can be identified (Kumar, 2020).

Feature-based theories take a de-compositional approach to meaning (Vigliocco & Vinson, 2012) whereby a concept is defined by a list of attributes or features, possibly together with their salience or typicality. Features are typically obtained from explicit human intuition, i.e., by asking speakers of a language to list the features they deem relevant to describe given words (Buchanan et al., 2019; McRae, 2005; Vinson & Vigliocco, 2008). Then, given two concepts, feature-based models define their semantic similarity based on the proportion of features they share. For example, consider

a simplified set consisting of just eight features and the overlap between the concepts of "dog" and "dolphin" (each concept possesses **bold** features, **blue** if also shared with the other):

$$dog = \begin{bmatrix} fly \\ \boldsymbol{fur} \\ \boldsymbol{mammal} \\ \boldsymbol{bark} \\ aquatic \\ sing \\ fins \\ \boldsymbol{smart} \end{bmatrix}, dolphin = \begin{bmatrix} fly \\ fur \\ \boldsymbol{mammal} \\ bark \\ \boldsymbol{aquatic} \\ sing \\ \boldsymbol{fins} \\ \boldsymbol{smart} \end{bmatrix} \rightarrow \frac{2}{8} = 0.25$$

Computational models of semantic processing (Cree et al., 1999) and learning (Saxe et al., 2019) can then be implemented on the basis of featural representations. Classical connectionist models from the parallel distributed processing group (PDP; Rumelhart et al., 1986; Hoffman et al., 2017) are a prime example of this implementation: units represent concepts (e.g., "dog"), features (e.g., "fly", "fur", "mammal"), and their relations (e.g., IS A, CAN, HAS), while edges represent the strength of association among units, which are trained to capture the appropriate relationships ("dog HAS fur", "dog IS A mammal").

Associative network models (also called *graphs*) shift the focus from the internal (featural) structure of concepts to their relationships. In a semantic network, nodes represent concepts, while edges (the links among nodes) specify their semantic relationships (Kumar, 2020). Associative networks are built from *associative relations*, the outcome of retrieval and free-association tasks (Collins and Loftus, 1975; De Deyne et al., 2019; Kumar, 2020). Then, processing accounts can be defined over graph representations; most notably, spreading activation (Collins and Loftus, 1975). The abstractions provided by graph structures and metrics can then be leveraged to study semantic phenomena at multiple levels (Kumar et al., 2021; Siew et al., 2019), i.e., "local" and "global" properties of graph connectivity (e.g., small-worldness; Petrou et al., 2010; Sigman and Cecchi, 2002).

The extent to which featural and associative information are distinct sources of semantic information is still debated (Kumar, 2020). However, both are themselves behavioral outcomes, which poses a problem of epistemological circularity for feature-based models and associative network models alike (i.e., behavioral phenomena – e.g., ratings, free-associations – are used to explain other behavioral phenomena – e.g., semantic priming; Jones et al., 2015; Westbury, 2016). Relatedly, both theories lack an account of how their representations came to be, that is, they lack an account of *learning* (Kumar, 2020). In this respect, a third class of models, distributional semantics models (DSMs), offer a compelling alternative.

1.1.2 Distributional semantics models

Distributional semantics models (DSMs; Landauer & Dumais, 1997; Mikolov, 2013) are computational models of meaning based on the *distributional hypothesis* (Harris, 1954; Firth, 1957), which states that words with similar meanings tend to occur in similar contexts (e.g., "fork" and "knife"). The availability of large textual corpora enables the operationalization of the distributional hypothesis based on co-occurrence information, whereby word meaning is first characterized as a distribution over contexts, which can be other words within a given distance, or context window, from the target word (Lund & Burgess, 1996), or larger contexts such as documents (Griffiths et al., 2007; Landauer & Dumais, 1997). For example, consider the sentence: "When it saw the cat, my dog started barking"; by defining a context window of size ±2, the word "dog" is assigned the neighbor words "cat", "my", "started" and "barking", which we can use to populate a word-by-context matrix where each entry codes for the number of times (or counts) the word co-occurred near all the other words in the model's vocabulary. By repeating the process for all instances of "dog" in a large corpus and updating the counts, we obtain a rich co-occurrence representation:

	a	abandon	ability	…	zoology	zoom	zucchini
dog	1205	6	2	…	26	0	0
house	3430	2	0	…	2	0	1

Figure 1.1: A simplified co-occurrence matrix for the words "dog" and "house". Each entry displays the number of times the word in the corresponding row occurred in the same context of the word in the corresponding column. Note that, because of its baseline frequency, the co-occurrence counts of the word "a" with "dog" and "house" are extremely high, despite "a" being uninformative of both meanings.

The semantic similarity of words is then defined as the overlap of their co-occurrence profiles. Note that this is different from the actual co-occurrence of the two words: words that occur together very rarely can nonetheless present very similar distributional profiles (e.g., near-perfect synonyms such as "overweight" and "corpulent"; Landauer & Dumais, 1997). However, raw co-occurrence counts conflate contingencies that are due to genuine semantic relationships with those that result from mere frequency. For example, the word "a" co-occurs very frequently with most words in the vocabulary, and therefore has an influence on their distributional profiles that is both outsized and aspecific. To counter this problem, DSMs apply weighting schemes based on word probabilities and information theoretic relations (e.g., pointwise mutual information; Church & Hanks, 1990). Then, the

dimensionality of these word-by-context representations is reduced via a series of techniques such as singular value decomposition (Martin & Barry, 2007). This step improves the performance of DSMs and is theoretically motivated by the goal of extracting distributional regularities that are general or *semantic*, and that underlie individual or *exemplar* word-by-context information (Günther et al., 2019). The result is a *semantic space* whose dimensions, while not directly interpretable, capture relevant, *latent* aspects of meaning, and where words are naturally represented as points defined by their coordinates, i.e., by *vectors* or *embeddings*. Consider again the meaning of "dog" and "house":

$$\overrightarrow{dog} = \begin{bmatrix} 0.15 \\ 0.02 \\ 0.11 \\ \vdots \\ 0.09 \end{bmatrix}, \overrightarrow{house} = \begin{bmatrix} 0.10 \\ 0.13 \\ 0.08 \\ \vdots \\ 0.19 \end{bmatrix}$$

Here, the semantic similarity of two words can be quantified by the proximity of their corresponding vectors. By far the most common metric of vector proximity is cosine similarity, or normalized dot product (figure 1.2):

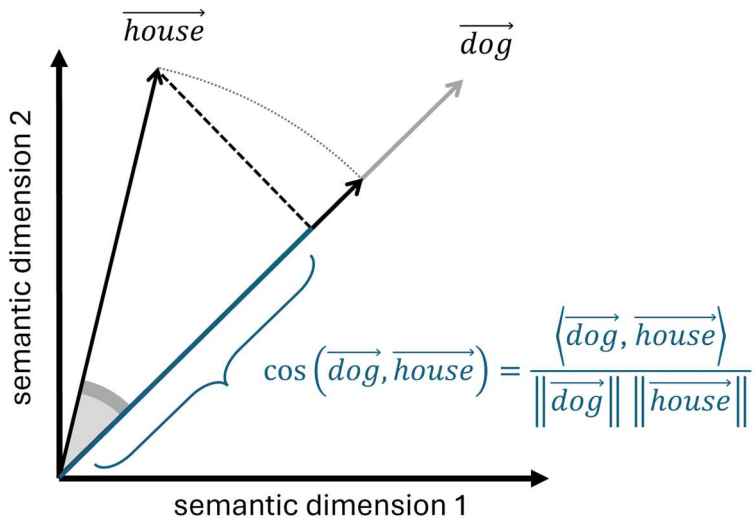

Figure 1.2: The "dog" and "house" embeddings in a simplified 2-dimensional semantic space, and their cosine similarity (i.e., the projection of a vector onto the other length normalization).

DSMs that follow the steps described above – modeling contexts from co-occurrence *counts* – are called *count-based models*. Learning in these models is typically non-incremental (but see BEAGLE; Jones & Mewhort, 2007), as it starts by assuming the availability of the entire co-occurrence experience, from which low-dimensional clusters are identified in an *error-free* fashion (Kumar, 2020). Examples of count-based models include hyperspace analogue to language (HAL; Lund & Burgess, 1996), latent semantic analysis (LSA; Landauer & Dumais, 1997), and topic models

(Griffiths et al., 2007). On the other hand, another family of DSMs, called *prediction-based models* (Baroni, Dinu, et al., 2014), includes models that acquire word-by-context representations via incremental learning mechanisms guided by expectation and error correction (for this reason, they are also called *error-driven* models; Kumar, 2020). A popular example (used throughout the present dissertation) is the word2vec model (Mikolov, 2013; Mikolov et al., 2013), a neural network with a single hidden layer. Word2vec has two versions: continuous bag-of-words (CBOW), where the model is trained to predict a word given its context, and skip-gram, where word2vec is given words as input and is trained to predict their context. Theoretically, error-driven models, which descend from the tradition of connectionism (Rumelhart et al., 1986; section 1.1), implement learning mechanisms equivalent to classical discriminative learning algorithms from animal learning (Rescorla & Wagner, 1972), and are thus funded on psychologically plausible principles (Hollis, 2017; Mandera et al., 2017). Therefore, contrary to feature-based and associative network models (section 1.1.1), DSMs present an account of semantic representation that is not grounded in semantic phenomena themselves and, at least in the case of error-driven models (and some count-based models), that is complemented by an account of how such semantic representations are learned.

Besides their theoretical appeal, DSMs have been able to account for a wide range of behavioral findings, such as synonym detection (Baroni, Dinu, et al., 2014; Landauer & Dumais, 1997; Mandera et al., 2017), free-association (Nematzadeh et al., 2017), concept categorization (Baroni, Dinu, et al., 2014), analogy (Levy & Goldberg, 2014; Mikolov, 2013), and feature ratings (Grand et al., 2022). Moreover, across many neuroimaging studies, DSMs proved to be powerful representational models of the brain correlates of human semantics (Carota et al., 2017; Huth et al., 2016; Mitchell et al., 2008; Pereira et al., 2018). Thus, on both theoretical and empirical grounds, DSMs can be considered psychologically plausible models of semantic representation and learning (Günther et al., 2019). Nonetheless, DSMs also have many limitations. First, DSMs represent the meaning of simple linguistic units (i.e., words and frequent multiword expressions), leaving out the issue of how these meanings are combined, e.g., to form a sentence. Relatedly, while conceptual representations are context-dependent (Yee & Thompson-Schill, 2016), DSMs' representations are "static" and "prototypical", namely unaffected by the broader sentential context in which concepts occur, and thus not accounting for the variability of different instances or exemplars of the same concept (e.g., the meaning of "house" should be different in the idiomatic expression: "A house of cards"). Second, contrary to humans, DSMs learn semantic representations from textual corpora and do not directly consider sensorimotor experience, which is crucial to reach a complete picture of semantics (Lake & Murphy, 2021). Different solutions have been proposed to address these limitations (see section 1.3.3 for the specific issue of conceptual combination in DSMs).

1.1.3 Large language models

Large language models (LLMs) are deep neural networks trained on large corpora for general purpose language understanding and generation. Motivated by the desire to focus on specific parts of a sentence for language translation, LLMs leverage the transformer's attention mechanisms (Vaswani et al., 2017) to learn the relations between different parts of a text sequence. More specifically, LLMs first decompose an input sequence – typically, a sentence – into tokens, namely subword units whose meaning, like DSMs, is described by high-dimensional embeddings (in this sense, LLMs are "grounded" on distributional representations). Unlike DSMs, however, LLMs' attention mechanism captures the dependencies among tokens, updating their embedding according to the broader context of the sequence, resulting in *contextualized* word embeddings (CWEs). Formally, token embeddings are projected to the lower dimensional spaces of queries Q, keys K, and values V, and the updated embedding of each token is defined by the sum of all values V weighted by the compatibility of the target token's query and all keys (figure 1.3):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where $\sqrt{d_k}$ is the dimensionality of the embeddings, and the $softmax$ function amplifies the differences in the key-query similarities $\frac{QK^T}{\sqrt{d_k}}$ and converts them into probabilities. Multiple attention heads are learned and applied in parallel, and their result is aggregated and passed to a subsequent feedforward neural network. The process is repeated through a series of layers until an output is reached, which varies among different LLMs and training objectives, e.g., reconstructing a masked-out portion of text sequences, or "masked-language modeling" (e.g., BERT; Devlin et al., 2018), or modeling the probability of upcoming tokens (e.g., GPT-2; Radford et al., 2019). Besides differences in architecture and objective, LLMs present an appealing alternative to DSMs in that they naturally incorporate token combination and context-dependence, albeit introducing considerable architectural complexity.

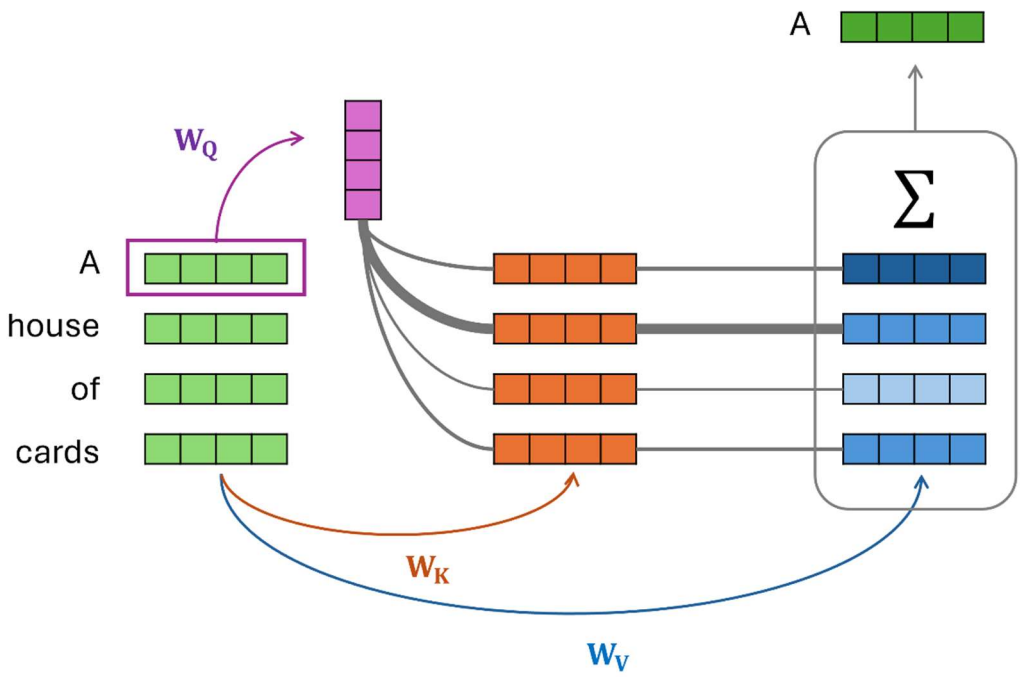

Figure 1.3: The self-attention mechanism operating on the example sentence: "A house of cards", tokenized into "A", "house", "of", "cards". To obtain the updated embedding of a token (here, "A"), the W_Q, W_K and W_V matrices generate the queries Q, keys K,and values V embeddings (colored purple, orange, and blue, respectively). Keys and queries are multiplied (QK^T) to obtain their similarity (gray lines). After applying dimensionality normalization and a $softmax$ transformation, these similarities define the weights used for the weighted sum of value vectors (here, darker shades of blue for larger weights), which defines the updated embedding.

1.1.4 Data-driven models beyond language

Because DSMs and LLMs alike learn semantic representations from language form alone, it has been argued that these models cannot, even in principle, learn meaning, which on some accounts is essentially referential (Bender & Koller, 2020; but see Piantadosi & Hill, 2022). A related, more empirically-oriented criticism concerns their alleged inability to capture extralinguistic information, which is seen as essential for a comprehensive account of conceptual representation (especially by embodied accounts of cognition; Barsalou, 2008; see section 1.3.4). On one hand, some efforts have been devoted to evaluate this claim, showing that despite being confined to patterns encoded in text, these are sufficient to capture (some) extralinguistic information (Gatti, Marelli, Vecchi, et al., 2022; Louwerse, 2011; Louwerse & Benesh, 2012). On the other hand, data-driven models were developed targeting extralinguistic information, most notably vision, either by developing multimodal models (Bruni & Baroni, 2014; Radford et al., 2021) or by focusing on alternative modalities (Günther et al., 2022; Lindsay, 2021; Wichmann & Geirhos, 2023).

Relevant to the present dissertation, Günther et al., (2022) proposed to represent high-level visual information by leveraging the semantic spaces learned by convolutional neural networks (CNNs) trained on image classification. CNNs are deep neural networks whose core architecture comprises a series of convolutional and fully connected layers (see figure 1.4). In input, CNNs receive an image, which can be represented as a tensor of RGB values. Convolutional layers apply convolution operations, namely filters that extract higher-level features from the image (or the preceding layer). These abstract features are then additionally transformed by fully-connected layers, which culminate in a classification layer where each unit or "neuron" corresponds to a concept from the set of classification categories (typically, the 1000 categories employed in the ImageNet Large Scale Visual Recognition Challenge; Russakovsky et al., 2015). Like DSMs semantic spaces, vision spaces represent images as high-dimensional vectors, and visual similarity is quantified by measures of vector proximity (such as cosine similarity). Unlike DSMs, vision spaces are based on *visual* features, whose abstraction depends on the specific layer (with more "abstract" and "semantic" information represented in deeper layers). The different weight given to semantic and visual properties results in different, complementary organizations of the same concepts in the two spaces. For example, figure 1.4 illustrates how "drill" and "screwdriver" are more similar in the space of a DSM since they are both tools, while "drill" and "gun" are more similar in the space of a CNN since they are perceptually closer. Note that, while concepts are identified with individual words in DSMs, multiple images can describe the same concept; these can be considered "exemplars" of that concept, and a summary operation – typically, their average – can be used to generate the corresponding prototype (Günther et al., 2022). In a large-scale evaluation, Günther et al., (2022, 2023) showed that the vision spaces provided by CNNs layers predict behavior across untimed (ratings of image similarity, visual word similarity, and image typicality) and timed tasks (priming and discrimination tasks).

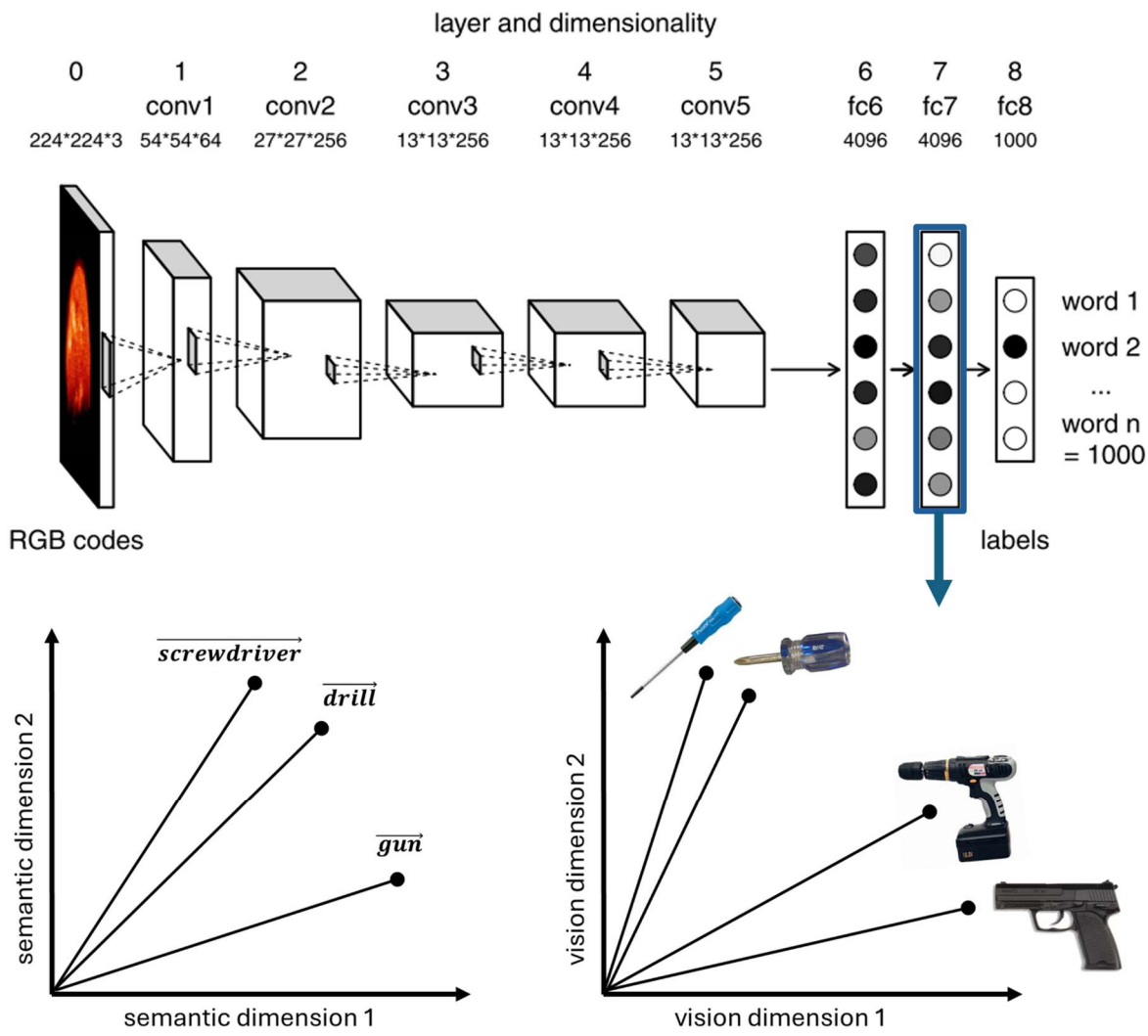

Figure 1.4: Simplified illustration of a vision space (adapted from Günther et al., 2022). (top) Illustration of the VGG-F convolutional neural network architecture (Chatfield et al., 2014). (bottom) Simplified 2-dimensional semantic spaces generated by a DSM (bottom left) and by the seventh layer of VGG-F (bottom right), representing the concepts of "drill", "gun", and "screwdriver". Moreover, because the same concept (here, "screwdriver") can be depicted by more than one image, a concept can be represented by multiple vision embeddings.

1.2 Conceptual combination

1.2.1 Theories of conceptual combination

Having discussed the representation of individual concepts in relation to main cognitive theories of semantics, I now turn to the central theme of this dissertation, i.e., how individual concepts are combined, a topic that has given rise to multiple theories of conceptual combination over the years (for a complete overview, see Ran & Duimering, 2009)..

The concept specialization model (Cohen & Murphy, 1984; Murphy, 1988; 2004) assumes that concepts are schemas defined by "slots" of abstract dimensions and "filler" values. According to this model, background knowledge guides first the selection of appropriate dimensions of the head concept, which are "filled" by the modifier concept (e.g., "house" specifies the "location" attribute of "dog" in "house dog"), and then a process of elaboration whereby the resulting combination is further specified and made more coherent by drawing on external background knowledge (e.g., a "house dog" is more likely to be well-behaved, clean, and small).

The dual-process model (Wisniewski, 1996) also assumes a schematic representation of concepts, but distinguishes between two types of interpretations – property-based and relation-based – that operate in parallel, resulting in combinations that can be divided into three classes: *property-mapping* (where a property of the modifier is assigned to the head; e.g., "robin hawk" is "a hawk with a red underbelly"), *relation-linking* (where the modifier and head concepts are related by a thematic relation; e.g., "taxi driver" is "a driver of taxis"), and *hybrid* or *conjunctive* (the combination is a cross between head and modifier – e.g., "robin hawk" is "a cross between robins and hawks" – or is *both* the head and modifier – e.g., "singer-songwriter" is someone that both sings and writes her songs). As Pezzelle & Marelli, (2020) pointed out, linguistic theorizing by Scalise & Bisetto, (2012) converged on an analogous, tripartite taxonomy based on the on the type of implicit syntactic relation that links constituents, dividing compounds into *subordinate* (the constituents are related by a complement relation), *attributive* (the modifier specifies an attribute of the head), and *coordinate* (the constituents are linked by the logical conjunction "and"). The division between "attributive" and "relational" combinations can be found in other models such as the interactive property attribution model (Estes & Glucksberg, 2000) and the retrieval–composition-analysis model (Prinz, 2002).

The CARIN (Gagné & Shoben, 1997) and RICE (Spalding, Gagné, Mullaly, & Ji, 2010) theories posit that all combinations are relational, and subsume attributive, relational and coordinate compounds in a general framework of semantic relations (i.e., attributive combinations are seen as a special type of semantic relation; e.g., "goldfish" could be described by the IS relation as "fish IS gold"). These theories propose that compound word processing involves the automatic decomposition of a compound (e.g., "snowman") into its constituents ("snow" and "man"), which are then connected by an implicit semantic relation selected among a set of competitors (FOR, BY, <u>MADE OF</u>, HAS, …) to obtain a plausible interpretation ("man MADE OF snow").

Finally, Costello & Keane, (2000) propose the constraint model of conceptual combination, based on pragmatic principles whereby an interpretation is selected based on the principles of plausibility (the interpretation is consistent with prior knowledge), diagnosticity (the combined

concept is best identified by its constituents) and informativeness (i.e., both constituents are necessary and sufficient to convey the desired information). Notably, Costello & Keane present a computational implementation of the constraint theory, the C^3 model, that assumes a schematic representation of concepts based on abstract features.

1.2.2 The neural bases of conceptual combination

Neuroimaging studies, by investigating the neural correlates of simple two-word phrases, have identified multiple regional and network specializations as the main substrates of conceptual combination. Of those, four regions of interest have been consistently observed (Coutanche et al., 2018)[1]: left anterior temporal lobe (LATL), the left and right angular gyri (LAG, RAG), and the left inferior frontal gyrus (LIFG).

In line with its role as a multimodal semantic hub (Lambon Ralph et al., 2010; Ralph et al., 2016), the LATL appears to support the integration of semantic information (Coutanche et al., 2018; Frankland & Greene, 2020). Evidence from MEG studies suggests that early LATL activation is modulated by the magnitude of conceptual specificity updating (Pylkkänen, 2019; Westerlund & Pylkkänen, 2014; Zhang & Pylkkänen, 2015; Ziegler & Pylkkänen, 2016), and that the LATL supports combinations based on extracting shared features, or *intersective conjunctions* (Poortman & Pylkkänen, 2016), with fMRI studies advancing analogous considerations (Baron & Osherson, 2011; Coutanche & Thompson-schill, 2015). Moreover, meaning combination in LATL has been characterized as largely independent of plausibility and syntactic composition (even though these factors might lead to additive effects; Parrish & Pylkkänen, 2022).

RAG and, most notably, LAG activation have been related to compound meaningfulness (Forgács et al., 2012; Graessner et al., 2021; Graves et al., 2010; Price et al., 2015; Price et al., 2016). While these findings would follow naturally from combinatorial processes in the AG, Ralph et al., (2016) and Humphreys et al., (2021) raised the possibility that, in line with its contribution to the default mode network, AG activation may simply index task difficulty (but see Boylan et al., 2017). Concerning AG computations, Coutanche et al., (2018) argue that AG support *relational combinations* (see Boylan et al., 2017), consistent with evidence suggesting that AG represents

[1] Besides these ROIs, other regions have been implicated in meaning combination, most notably the ventromedial prefrontal cortex (vmPFC), posterior cingulate cortex (PCC), left middle-superior temporal cortex (lmSTS) (Frankland & Greene, 2020; Pylkkänen, 2019), which, together with the LATL, AG, and LIFG, mostly lie within the broader semantic network (Binder et al., 2009; Jackson, 2021) and the default mode network (Frankland & Greene, 2020).

semantic relations (Wang et al., 2020; Zhang et al., 2022), and thematic relations more broadly (Schwartz et al., 2011).

Finally, LIFG (specifically *pars triangularis* and *pars orbitalis*) activation has been linked to *decreased* compound meaningfulness (Forgács et al., 2012; Graessner, Zaccarella, Friederici, et al., 2021; Graves et al., 2010; Molinaro et al., 2015) and increased feature uncertainty (Solomon & Thompson-Schill, 2020), suggesting that this region may be recruited to process more effortful combinations, possibly mediated by its known involvement in feature selection (Coutanche et al., 2018), and consistent with its role within the semantic control network (Jackson, 2021). As with the AG, it has been claimed that LIFG involvement might reflect task difficulty rather than actual combinatorial processing.

1.2.3 The role of morphological structure

Similarly to inflection and derivation, extensive evidence shows that compound word processing is influenced by lexicosemantic properties of constituent words, most notably their frequency and positional family size (i.e., the number of compounds in which a word appears as a modifier or head constituent, or "family members"; Gagné, Spalding & Schmidtke, 2019; Schmidtke et al., 2021; Baayen, Kuperman, & Bertram, 2010), and secondary family size (i.e., the positional family sizes of a constituent's family members; Baayen, 2010). Moreover, constituent position has been shown to interact with effects of compound semantics ("relational dominance", Gagné, 2001; Gagné & Shoben, 1997), family frequency (Gagné & Spalding, 2009) and headedness (Marelli & Luzzatti, 2012). Thus, rather than retaining their "free word" meaning, constituents tend to form morphological families that share systematic semantic tendencies (e.g., "-berry" in "strawberry", "raspberry", "blueberry", etc., or the "bat-" modifier in "Batman", "batmobile", "batplane", etc.), and new compounds tend to be coined by re-using constituents from productive morphological families (Libben, 2014). These findings have led to the formulation of the *morphological transcendence hypothesis* (Libben, 2014), which claims that, as a result of language exposure, the original meaning of a constituent (as an independent lexeme) progressively drifts towards a "transcended" form linked to its positional role within compounds. For example, with the recurrent use of the word "bag" within various compound words, the positionally bound meanings of "bag-" and "-bag" are developed, which can be (and often are) different from the original meaning of "bag" and/or from each other (e.g., in "bagpipe" and "windbag").

1.2.4 The role of sensorimotor information

Theories of conceptual combination critically depend on broader assumptions about what concepts are, i.e., on theories of conceptual knowledge. Some of the existing theories of conceptual combination are not committed to specific types of semantic representations (e.g., CARIN), while others take an "amodal" view, whereby the unit of representation are symbols, or slots, corresponding to abstract features (e.g., dual-process theory, concept specification theory). Instead, according to theories of embodied or grounded cognition (Barsalou, 2008; Glenberg et al., 2013), sensorimotor experience is not only essential to build conceptual representations, but conceptual processing itself entails the partial re-instantiation (i.e., a conscious or unconscious form of re-experience or *simulation*) of such sensorimotor information, often in close dependence with contextual information and goals (hence, conceptualizations are *situated*; Barsalou, 2009, 2020). For example, the concept of "apple" will comprise features that are systematically experienced by interacting with apples (e.g., shape, color, taste, weight), which are partially re-instantiated when using the "apple" concept in a context-dependent fashion (e.g., throwing an apple might involve the re-instantiation of weight and shape properties, but not those of color or taste). According to these theories, conceptual combination will also depend on (or be constituted by) sensorimotor experience and simulation. Indeed, Wu & Barsalou (2009) show that the generation of compound properties depended on perceptual characteristics such as occlusion, even when perceptual properties are not explicitly elicited, suggesting that conceptual combination relies on automatic sensorimotor simulations. Günther, Petilli, et al. (2020) demonstrate effects of combinatorial visual information over and above a combinatorial language baseline (see section 1.3.2) on large-scale datasets of naming, lexical decision and timed sensibility judgements, suggestive of automatic influence of perceptual information on the on-line processing of compounds. In the context of grounded cognition, the Embodied Conceptual Combination (ECCo) theory (Lynott & Connell, 2010) advances a specific account of conceptual combination. Specifically, the ECCo model posits that conceptual combination depends on situated simulations, whereby sensorimotor features of the constituents are re-instantiated and meshed according to their compatibility or "affordance". At the same time, ECCo also holds that distributional linguistic patterns provide complementary cues to the simulation system, providing coarse-grained information jointly and constraining the process of combination. The two systems are theorized to interact dynamically in time, with a tendency for faster engagement of the linguistic system, and with their relative importance dependent on the specifics of constituents, context, and task demands. Connell & Lynott (2011) provide initial evidence for the ECCo model on timed-sensibility judgements.

1.2.5 The interplay between storage and computation

Familiar compounds can be seen as units at the crossroads between words and sentences (Libben, 2010). On the one hand, their meaning could be simply stored and retrieved from long-term memory, similarly to morphologically simple words whose meaning cannot be computed but has to be memorized. On the other hand, because the meaning of most compounds can be predicted, at least in part, from their constituents (i.e., most compound words are *semantically transparent*; Gagné et al., 2019), it is often possible to understand a compound by active combination, similarly to how the meaning of most sentences must be computed rather than memorized. This raises the issue of whether and to what extent compounding relies on retrieval from memory or on online computations.Theoretical arguments and empirical evidence suggest that compound word processing leverages both retrieval from long-term memory (if the word is familiar) as well as automatic constituent combination to harvest candidate compound meanings, maximizing the opportunities of successful comprehension (Libben, 2014; Libben et al., 1999; Semenza & Mondini, 2010). Indeed, from a theoretical perspective, combination is clearly required when a compound is first encountered. Because it cannot be known in advance whether a compound word is novel and/or transparent, in order to avoid pervasive sources of error (e.g., incorrect parsing) composition should be automatically attempted. This consideration is of prime relevance since compounding is a productive phenomenon of word formation in some languages, which implies that speakers routinely encounter novel compounds that must be interpreted on the spot (Libben, 2010), subjecting pragmatic constraints on novel combinations (see Costello & Keane, 2000). Second, many empirical findings are difficult to reconcile with theories of compound processing that rely only on static connections among lexicalized items (Chamberlain et al., 2019; El-Bialy et al., 2013; Ji et al., 2011). In this context, many studies relied on the effects of semantic transparency on behavioral measures. Semantic transparency, although defined and operationalized differently across studies, generally refers to the extent to which the meaning of a constituent (or both) contributes to the meaning of the compound (Auch et al., 2020), spanning a continuum from full opacity (e.g., “hogwash”) to full transparency (e.g., “snowball”). Transparent compounds are typically processed faster than opaque compounds and monomorphemic controls (Chamberlain et al., 2019; Ji et al., 2011; Kim et al., 2019; but see Schmidtke, Van Dyke, et al., 2018 for an overview of inconsistent findings). Notably, and contrary to what models of pure storage-retrieval would predict, semantic transparency interacts with other factors such as the position of a constituent and the headedness of the compound (Marelli & Luzzatti, 2012), the frequency of the modifier (Ji et al., 2011) and the head (Marelli & Luzzatti, 2012), the transparency of the other

constituent (El-Bialy et al., 2013) and the particular dimension of transparency captured by the other constituent (Auch et al., 2020). While the strategy of opportunity maximization described above concerns the comparison of "storage" and "compositional" routes, when taken to its logical conclusion, it implies that different *types* of composition should also be automatically attempted. From a theoretical standpoint, not only one doesn't know in advance whether a compound is novel, and hence necessitates active combination; one also doesn't know which *type* of combination is going to be appropriate. Indeed, the CARIN and RICE theories (Gagné & Shoben, 1997; Spalding, Gagné, Mullaly, & Ji, 2010) claim that interpretation results from a process of competition among multiple candidate semantic relations. Here, empirical evidence supporting the CARIN and RICE theories does suggest that different combination types, in the form of different relational interpretations, compete to be selected (Benjamin & Schmidtke, 2023; Schmidtke et al., 2016; Schmidtke, Gagné, Kuperman, & Tucker, 2018).

1.3 Computational models of conceptual combination

1.3.1 Motivating computational implementations of conceptual combination

As described in the above sections, theoretical accounts of conceptual combination posit that multiple representations *of the same concept* are developed or generated in parallel to aid successful comprehension: the meaning of a constituent as a standalone word is distinguished from its modifier- or head-dependent "morphologically transcended" meaning (Libben, 2014); concepts are characterized by distributional language information and by "simulated" sensorimotor features (Lynott & Connell, 2010); familiar compounds are both retrievable from long-term memory as "lexicalized" representations and computable on the fly via conceptual combination (Libben, 2010), while the latter itself can be accomplished in alternative ways, possibly supported by different brain regions (Boylan et al., 2017; Coutanche et al., 2018) and possibly carried out simultaneously (Benjamin & Schmidtke, 2023; Schmidtke, Gagné, Kuperman, & Tucker, 2018).

As stated at the beginning of this introduction, one of the main reasons why conceptual combination is interesting to study is that a lot of information required to make sense of it is implicit. For the same reason, however, some hypotheses can be challenging to test. Indeed, while the morphological transcendence hypothesis is grounded on various observations of the role-dependent behavior of constituent words (Libben, 2014), the construct of a transcended morphological unit is problematic to operationalize – due to its essentially latent nature – and thus problematic to test.

Likewise, the ECCo model posits sensorimotor and language-based representations for the same concepts, but lacks an explicit characterization of these two sources of information, leaving many of its predictions challenging to evaluate. Yet again, in order to test whether retrieval and combination (as well as different kinds of combination) are simultaneously attempted, it is helpful to represent the same combinations in alternative ways. While it is possible to address some of these questions with measures that should be correlated with the target constructs (e.g., frequency; Baayen et al., 2010), those are only indirect effects, i.e., cannot be identified with the construct. For example, many neuroimaging studies of semantic combination have largely focused on isolating brain activity related to high-level factors (e.g., comparing plausible and non-plausible combinations), thus overlooking how the underlying representations are combined, which is what conceptual combination *is*: a transformation of meaning-carrying representations. In this dissertation, I propose to address these issues by means of computational models. By defining both the representations of individual concepts and the functions that combine them, computational models provide a window into representations and processes that cannot be directly examined. In turn, because accounting for conceptual combination is a desideratum that any (computational) theory of human semantics must meet (Lake & Murphy, 2021), the empirical evidence overviewed above offers a challenging testing ground for models of conceptual combination, and of human semantics more generally.

Focusing on cognitive models, while those from the associative network tradition lack an account of how constituent representations are combined (Kumar et al., 2021; but see Kenett & Thompson-Schill, 2020), models based on feature combination have been proposed, such as Costello & Keane, (2000) C^3 model. These models, however, inherit the virtues and pitfalls of the representations they are based upon. Thus, for the same reasons advanced in favor of DSMs over feature-based and associative network models (section 1.1.1), I focus on compositional models based on distributional semantics.

1.3.2 Compositional distributional semantics models

The representational unit of DSMs consists mostly of single (content) words and frequent multiword expressions (e.g., “red herring”), leaving out larger, compositional expressions. Here, *compositionality* refers to the idea that the meaning of a complex expression is a function of the meaning of its constituents and of the way they are combined. Thus, scholars sought to develop compositional models (functions) operating over distributional semantics representations, or *compositional* distributional semantics models (cDSMs; Baroni, 2013; Baroni et al., 2014). To this

end, Mitchell & Lapata, (2010) explored simple cDSMs based on arithmetic operations among vectors. Two examples – taken on in Chapter 4 – are the additive model, which defines composition as simple vector addition (figure 1.5):

$$\vec{c} = \vec{u} + \vec{v}$$

And the multiplicative model, where composition is defined as the elementwise multiplication of constituent vectors (i.e., the ith element of $\vec{c}$, c_i, is given by the product of the corresponding elements of $\vec{u}$ and $\vec{v}$, $u_i v_i$):

$$\vec{c} = \vec{u} \odot \vec{v}$$

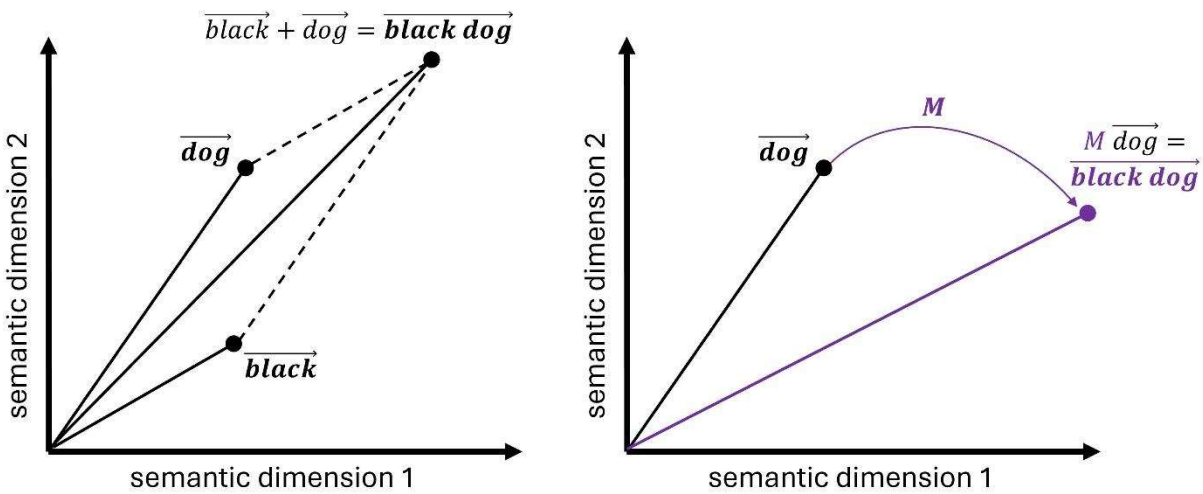

Figure 1.5: Geometric interpretation of cDSMs. Words are represented by vectors spanning a semantic space (shown here is the simplified case of vectors from a 2-dimensional semantic space, although cDSMs can be applied to vectors spanning semantic spaces of an arbitrary number of dimensions). (left) Geometric representation of the additive model. (right) Following a functional approach to semantic combination, the adjective "black" is formalized by the matrix *M*; geometrically, matrix application re-locates vectors in semantic space according to the properties typically brought by the adjective (estimated at training).

On the other hand, a different research program took inspiration from formal semantics' treatment of compositionality as function application (Baroni, Bernardi, et al., 2014; Baroni & Zamparelli, 2010). Because, in the language of DSMs, the meaning of words is described by vectors, functions were instead defined by higher-dimensional linear algebraic structures, i.e., tensors. In this context, adjectives have been defined by matrices (2-dimensional tensors), namely functions that take (noun) vectors as their arguments and output another, linearly transformed, (adjective-noun) vectors (Baroni & Zamparelli, 2010; figure 1.5):

$$\vec{c} = M\vec{v}$$

Where M is a $d \times d$ matrix that linearly transforms the d-dimensional vector $\vec{v}$. Functions such as M have free parameters that must be estimated, which is typically done by training a linear regression to predict the vectors of frequent compositional expressions from those of their constituents (Baroni, 2013; Dinu et al., 2013; Guevara, 2010). For example, given the tuples ("black hole", "hole"), ("black sheep", "sheep"), …, ("black coat", "coat"), the parameters of the matrix of "black", M, will be estimated with the objective of obtaining $\overrightarrow{black\ hole} = M(\overrightarrow{hole})$, $\overrightarrow{black\ sheep} = M(\overrightarrow{sheep})$, …, $\overrightarrow{black\ coat} = M(\overrightarrow{coat})$. Here, the high frequency of combinatorial expressions such as "black hole" allows for their vectorial representation in semantic space as single tokens, hence usable for training. After parameter estimation, the matrix M can be used to transform new nouns, even when the corresponding adjective-noun phrase does not occur in a corpus. Compared to arithmetic operations, matrix multiplication is more complex in that the application of the same matrix to different vectors can have very different consequences. For example, the matrix $M = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$ does not change vector $\vec{v} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$, since $\vec{c} = \begin{bmatrix} 0 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \cdot 0 + 0 \cdot 3 \\ -1 \cdot 0 + 1 \cdot 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \end{bmatrix} = M\vec{v}$; however, the same matrix leads to large changes of vector $\vec{u} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$, since $\vec{c} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \cdot 3 + 0 \cdot 3 \\ -1 \cdot 3 + 1 \cdot 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \end{bmatrix} = M\vec{v}$. This is a desirable property for conceptual combinations, since we know (section 1.2.1) that constituent meanings interact when combined. Following the example above, if $\vec{v}$ = "coal" and $\vec{u}$ = "eye", the matrix M would have correctly captured the fact that "black coal" is almost identical to "coal", while a "black eye" is not quite a mere "eye". Exploring different cDSMs and focusing on adjective-noun and noun-noun combinations, computational linguistic evaluations demonstrated that cDSMs predict the semantic similarity of combination pairs (Mitchell & Lapata, 2010), their paraphrases (Dinu et al., 2013b), and their plausibility (Vecchi et al., 2017). Beyond their performance, some cDSMs have been proposed as cognitive models of complex word semantics (Marelli et al., 2017; Marelli & Baroni, 2015).

In this context, based on a linear regression model developed by Guevara, (2010), Marelli et al., (2017) introduced the Compounding as Abstract Operations in Semantic Space (CAOSS) model, a compositional model with the following architecture (figure 1.6):

$$\vec{c} = M\vec{u} + H\vec{v}$$

Where $\vec{u}$ and $\vec{v}$ are, respectively, the modifier and head vectors, and M and H are matrices estimated by regressing $\vec{u}$ and $\vec{v}$ to predict $\vec{c}$ (figure 1.6c). Any given word vector multiplied by M (H) will inherit the typical semantic behavior of modifier (head) constituents. Last, the vectors resulting from these transformations are added to obtain $\vec{c}$, the compositional compound representation. Note that,

unlike the additive and multiplicative models, CAOSS is not commutative ($\overrightarrow{housedog} \neq \overrightarrow{doghouse}$). To illustrate the procedure, figure 1.6 shows the application of CAOSS to derive the compositional representation of $\overrightarrow{doghouse}$ from its constituents, $\overrightarrow{dog}$ and $\overrightarrow{house}$. The application of CAOSS to input word pairs results enables the compositional operationalization of different constructs.

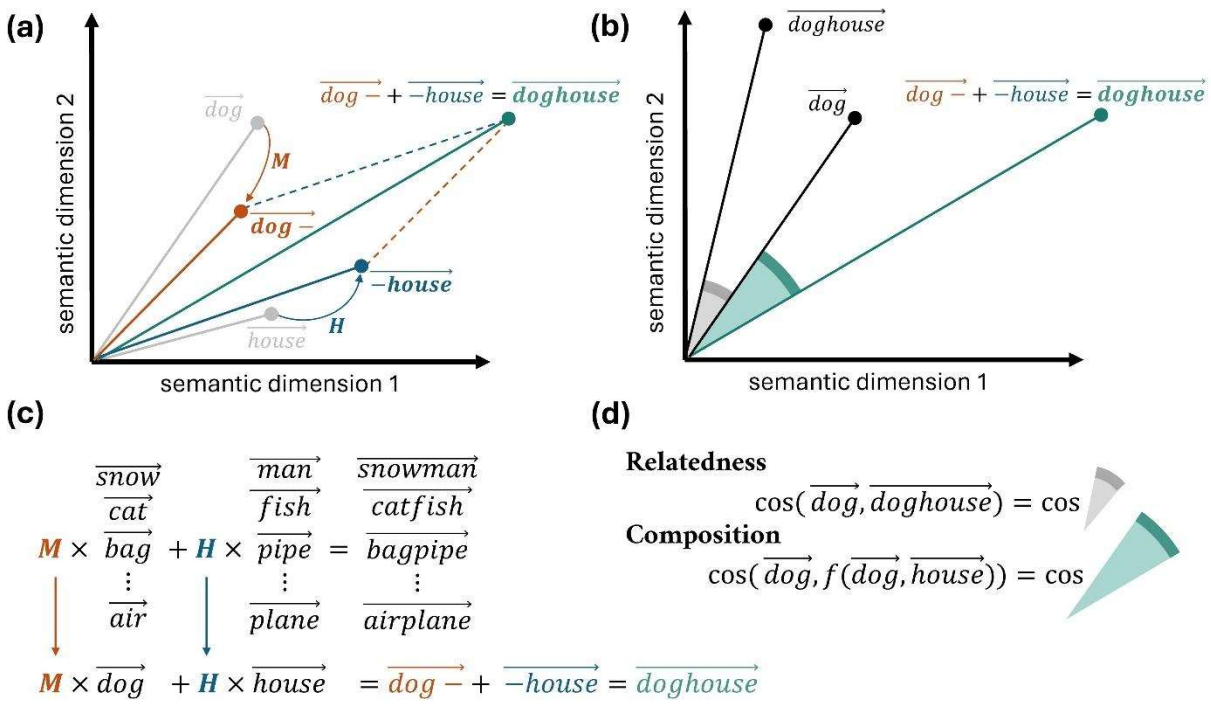

Figure 1.6: Compounding as Abstract Operation in Semantic Space (CAOSS) model architecture. (a) Geometric interpretation of the CAOSS to derive the compositional representation of "doghouse". The word vector of "dog" is multiplied by the modifier matrix *M*, obtaining the as-modifier representation "dog–". Analogously, the representation of "house" is multiplied by the head matrix *H*, obtaining the as-head representation "–house". Finally, the two as-constituent representations are summed together to obtain the compositional compound representation "doghouse". (b) The same compound can be described by "lexicalized" (DSMs-based) or compositional (cDSMs-based) compound representation. (c) A schematized sequence of CAOSS learning and inference: A list of lexicalized representations of (modifier, head, compound) triplets are used to train the model to obtain the parameters (weights) of the *M* and *H* matrices; at inference, *M* and *H* can be applied to a new compound. (d) Semantic transparency can be operationalized based on the similarity of a constituent and the "lexicalized" compound (gray angle) or compositional compound (green angle).

First, CAOSS replicated previous findings within the framework of the CARIN and RICE theories (Gagné & Shoben, 1997; Spalding et al., 2010; Marelli et al., 2017). Specifically, CAOSS replicated the relational priming effect, i.e., the finding that prime and target compound words that share the modifier and the relational interpretation are processed faster (e.g., "snowball" – "snowman", which share the modifier "snow" and the relation MADE OF; Gagné & Spalding, 2009)

and the relational dominance effect, i.e., the finding that atypical usage of the modifier leads to increased processing times (Gagné & Shoben, 1997; Gagné & Spalding, 2004). While previous works relied on handcrafted taxonomies of relational information (i.e., explicit human judgements of compound interpretation), CAOSS demonstrated that these effects can emerge from distributional patterns of language usage in a data-driven framework. Although CAOSS vectors are not directly interpretable, they implicitly encode human intuitions about (often multiple and conflicting) interpretations: Günther & Marelli, (2021b) demonstrated that a simple linear regression could successfully recover human intuitions on the interpretation of attested and novel compounds.

Second, thanks to the intermediate representations of words in their constituent-dependent role, CAOSS naturally provides an explicit characterization of as-constituent, or "morphologically transcended" units (figure 1.6a). Leveraging these latent representations, Günther & Marelli, (2021) were able to test differences between modifier and head constituents predicted by the morphological transcendence hypothesis in Libben, (2014).

Finally, while previous studies defined transparency to be the cosine similarity between the whole-word representation of compounds and their constituents, Günther & Marelli, (2018) argued that such measures consider words as independent units, neglecting the compositional relation that links them. Thus, Günther & Marelli, (2018) introduced a novel operationalization of semantic transparency based on explicitly combinatorial relations, namely the cosine similarity between whole-word constituents and the *compositional* compound representation (figure 1.6d). Indeed, Günther & Marelli, (2018) found that the newly introduced measure of semantic transparency, but not semantic relatedness, predicted the response times of lexical decision experiments, a finding later replicated with German compounds (Günther, Marelli, et al., 2020) and in a multimodal instantiation of CAOSS (i.e., when the model was augmented with vectors encoding visual information; Günther, Petilli, et al., 2020; see section 1.1.4). Moreover, measures of the contribution of constituent meaning to the combination successfully predicted response times to novel English compounds (Günther & Marelli, 2020).

1.3.3 Conceptual combination in large language models

Despite their successes, cDSMs inherit most of the limitations of the DSMs on which they are based (see section 1.1.2). While the evidence described above suggests that they can be adequate compositional models, cDSMs are problematic to scale up to longer sequences beyond simple phrases and conceptual combinations. Indeed, simple operations like those proposed by Mitchell & Lapata,

(2010) lead to an "information bottleneck" where all information must be somehow contained in the arithmetic combination of word vectors. Instead, formalizing composition via higher-dimensional structures presents other problems, among which are the insufficient data to train tensors whose parameters scale exponentially with compositional complexity (d^k for k-dimensional functions in a d-dimensional vector space; Stewart & Eliasmith, 2011), and the fact that tensors of different dimensionality, despite possibly carrying similar meanings, are incommensurable (e.g., "length" and "long" would be a (noun) vector and a matrix (adjective), respectively; Baroni, Bernardi, et al., 2014). These limitations can be understood as the result of cDSMs bottom-up approach to semantics, whereby complex meaning is *constructed* from simpler representations. Instead, large language models (LLMs), while grounded in the same distributional representations that constitute DSMs, adopt a complementary, top-down approach, where the primary representational unit is the sentence (Pavlick, 2022). Because the basic units of representation in LLMs are subword tokens, simple words and compound words are qualitatively identical, the latter simply being more amenable to subword tokenization (e.g., "snow" corresponds to the single token embedding $\overrightarrow{snow}$, while "snowman" to the pooled – typically, averaged – token embeddings $\overrightarrow{snow}$ and $\overrightarrow{\#\#man}$). Similarly, novel compounds such as "snowcastle" can be readily described by their subword embeddings ($\overrightarrow{snow}$, $\overrightarrow{\#\#castle}$). Thus, LLMs do not require dedicated operations to combine constituents; rather, conceptual combination is carried out by transformers' attention and is thus subsumed under a general-purpose combinatorial mechanism (see section 1.1.3 and figure 1.7). Moreover, because attention layers transform embeddings based on neighboring tokens, the same conceptual combination will be represented differently depending on the context in which it occurs. Finally, recent investigations of LLMs' representations of compound words (Buijtelaar & Pezzelle, 2023; Miletic & Schulte, 2023) and of words more broadly (Cassani et al., 2023) have led to encouraging results, suggesting that LLMs might offer an alternative to cDSMs as computational models of conceptual combination.

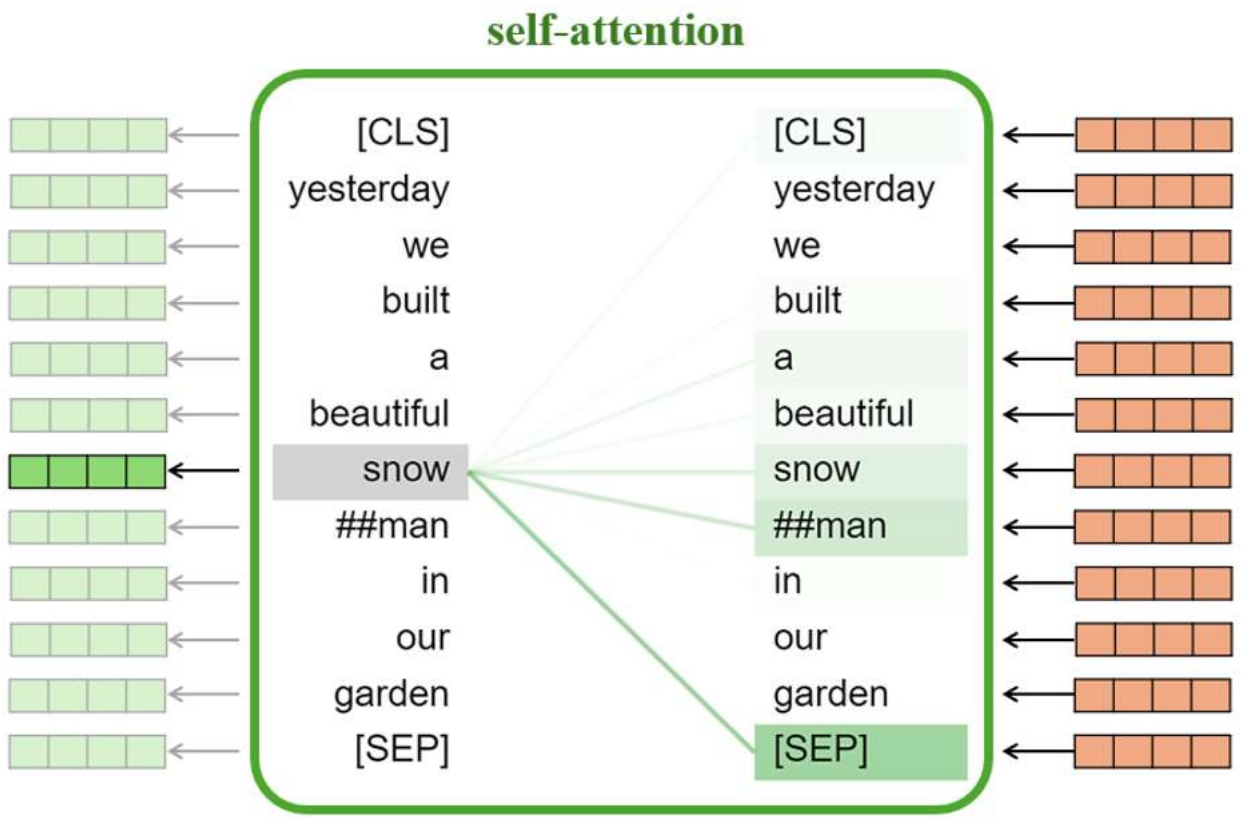

Figure 1.7: A simplified illustration of the attention mechanism operating over a compound word in a self-attention block from the BERT-base-uncased model (attention weight visualization using BertViz; Vig, 2019). The sentence "Yesterday we built a beautiful snowman in our garden" is decomposed into tokens (the [CLS] and [SEP] are special tokens indicating sequence beginning and separation, respectively). The compound can be also segmented into a variable number of tokens (here, "snowman" is decomposed into "snow" and "##man"). During a self-attention block, each compound token (here, "snow") attends to all others, including those that do not belong to the compound itself, with a degree of strength (or attention weight, shown by green edges) proportional to their contribution to the token's updating (i.e., the new embedding of "snow").

1.4 The current dissertation

Connecting the desiderata outlined in section 1.3.1 to the computational models overviewed in section 1.3, the following chapters are organized as follows:

i. In Chapter 2 CAOSS representations are used to test a claim of morphological transcendence, namely that "transcended" representations are learned after repeated encounters with compound words (Libben, 2014), which implies that their influence on behavior should depend on the frequency of such encounters. To do so, a large eye movement data from a large compound word reading dataset (Schmidtke et al., 2021) is re-analysed and, assuming "transcended" representations can be operationalized by CAOSS as-constituent

representations (following Günther & Marelli, 2021a), measures based on these intermediate representations are tested in interaction with measures of reading experience.

ii. Chapter 3 focuses on the interplay between language distributional information and visual information (Lynott & Connell, 2010) by formalizing the former with the CAOSS model (Marelli et al., 2017) and the latter with CNNs vision spaces (Günther et al., 2022). The contribution of the former is formalized by a language-to-vision mapping, which is leveraged to predict human behavior in response to synthetic images of novel conceptual combinations generated with a text-to-image model (Betker et al., 2023).

iii. Chapter 4 consists of a re-analysis of fMRI data aggregated from four published studies, all employing two-word combinations but differing in task requirements. The objective is to leverage what we know of cDSMs to test hypotheses of the type of conceptual combination carried out in target brain regions of interest (Coutanche et al., 2018). To do so, different types of semantic combination are formalized with three cDSMs, namely the additive, multiplicative and CAOSS models (Marelli et al., 2017; Mitchell & Lapata, 2010), and representational similarity analysis (Kriegeskorte et al., 2008) is used to relate their representations to those of the target ROIs. "Compositional" and "lexicalized" representations are available for the same stimulus, which allows to test the hypothesis that both processes are conducted in parallel.

iv. Chapter 5 explores the ability of cDSMs (Marelli et al., 2017; Mitchell & Lapata, 2010) and two LLMs, namely BERT-base (Devlin et al., 2019) and Llama-2-13b (Touvron et al., 2023), to represent the implicit relational interpretations of existing and novel compound words as defined by possible relation task judgements collected by Benjamin & Schmidtke, (2023), Schmidtke, Gagné, Kuperman, & Spalding, (2018) and Schmidtke, Gagné, Kuperman, & Tucker, (2018).

Chapter 2

Semantic transparency and reading experience

2.1 Introduction

A considerable body of studies shows that humans are sensitive to statistical regularities of their environment, and compounding is no exception. Both *individual-level experience*, (i.e., individual differences in language proficiency such as spelling and vocabulary), and *item-level experience* (item-specific measures such as word frequency and family size) influence compound word processing. This influence is typically beneficial: more frequent words, dominant interpretations, and greater participants' exposure to print lead to faster recognition times (Falkauskas & Kuperman, 2015; Marelli & Luzzatti, 2012), reading times (Kuperman et al., 2009; Marelli & Luzzatti, 2012; Schmidtke et al., 2021; Schmidtke, Van Dyke, et al., 2018), and more precise and flexible interpretations (Schmidtke et al., 2018). Individual- and item-level experience modulate the effect of form (Beyersmann et al., 2018; Falkauskas & Kuperman, 2015), lexical (Baayen et al., 2007, 2010; Schmidtke et al., 2021) and semantic (Schmidtke, Gagné, et al., 2018; Schmidtke, Van Dyke, et al., 2018) compound properties. Theoretical accounts explain these findings in terms of the progressive entrenchment of orthographic, phonological, syntactic, and semantic information, i.e., the simultaneous increase in flexibility and precision that results from their prolonged usage (Perfetti, 2007; Schmid, 2007). Within the domain of compounding, the *morphological transcendence hypothesis* (Libben, 2014) advances analogous considerations, namely that as a result of language exposure, the original, free-word meaning of a constituent progressively drifts towards a "transcended" form linked to its positional role.

The representational change posited by these theories should impact any compound word properties that draw on semantics, a key example of which is semantic transparency. Here, transparent

---

This work was presented at the 2022 edition of the "Italian Association of Cognitive Sciences" conference (Università di Trento, Rovereto, Italy) as the following contribution: "Ciapparelli, M., Günther, F., & Marelli, M. (2023). Semantic transparency in compound word reading: A computational investigation".

compounds are usually processed faster than opaque compounds and monomorphemic controls (Chamberlain et al., 2019; Ji et al., 2011; Kim et al., 2019). However, while the effect of transparency has been found to be modulated by multiple factors (Auch et al., 2020; El-Bialy et al., 2013; Ji et al., 2011; Marelli & Luzzatti, 2012), evidence for interaction effects with compound frequency is almost absent, despite the claim that the latter captures combinatorial knowledge (Baayen et al., 2007). Schmidtke, Van Dyke, et al., (2018) explain the conflicting evidence of transparency effects by considering two opposing drives. First, transparency *facilitates* processing due to the increased preactivation of the target compound resulting from the activation of (semantically similar) constituents. Second, according to the framework of the Naïve Discriminative Reader (Baayen et al., 2011), transparency *inhibits* processing due to the increased effort to discriminate the meaning of similar units. Only the latter encompasses learning, positing that learners continuously update their representations to discriminate better. Indeed, Schmidtke, Van Dyke, et al., (2018) show that semantic transparency was more beneficial for more experienced readers.

In this chapter, I propose that the impact of experience on semantic transparency can be modeled by employing the Compounding as Abstract Operations in Semantic Space (CAOSS; Marelli et al., 2017) model:

$$\vec{c} = M\vec{u} + H\vec{v}$$

Where $\vec{u}$ and $\vec{v}$ are, respectively, the modifier and head vectors. The modifier (M) and head (H) matrices are estimated by solving the closed-form solution of the least-squares linear regression problem (Dinu et al., 2013), which is the equilibrium state of a Rescorla-Wagner discriminative learning model (Evert, 2015; Gallistel, 2003; Hoppe et al., 2021). Therefore, the application of CAOSS matrices can be interpreted as a discrimination process which weights constituent words' semantic features according to how informative they are for predicting compound word features. Thanks to these operations, CAOSS enables a versatile operationalization of psycholinguistics constructs (figure 2.1).

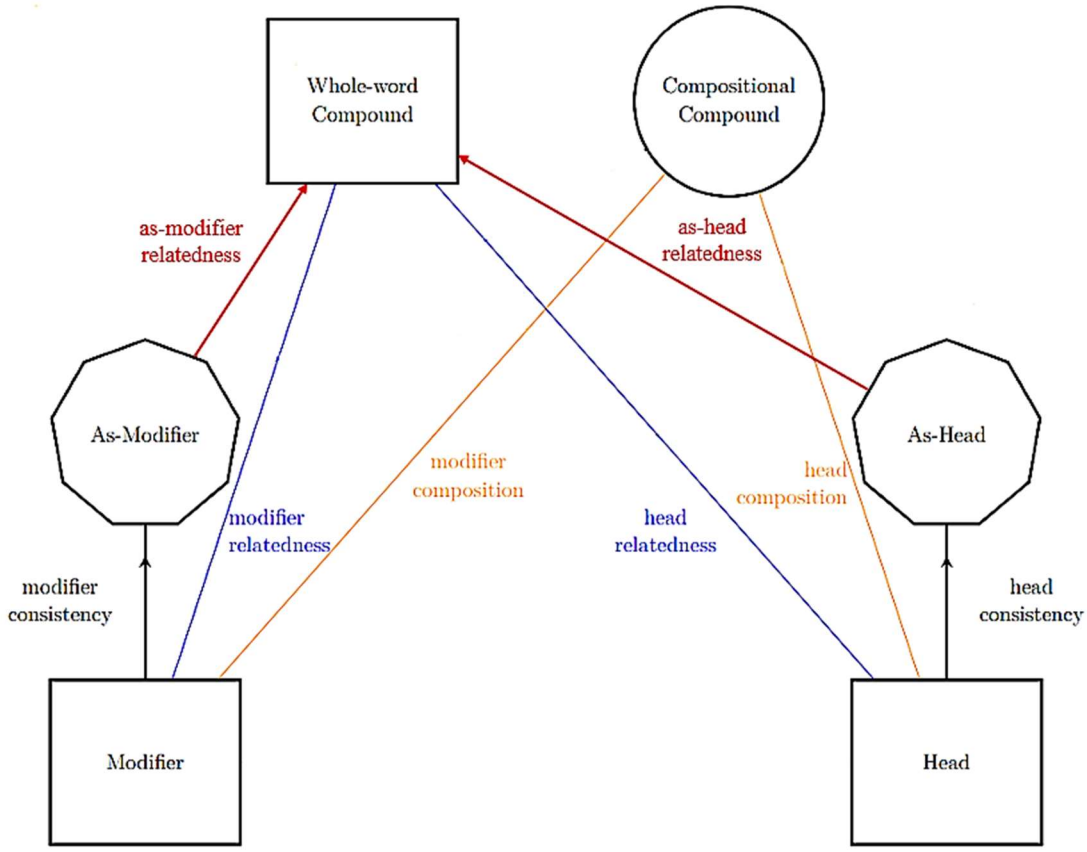

Figure 2.1: The complete set of meaning representations involved in the model: Lexicalized meanings are conceptualized as separate entries in semantic memory (squares), compositional meanings resulting from CAOSS (circle and polygons). The lines and arrows represent the semantic relations that connect meanings (quantified by cosine similarity): relatedness, consistency, composition, and as-constituent relatedness.

First, CAOSS offers an explicit characterization of the role-dependent representation of constituent words. Indeed, by exploiting these latent representations, Günther & Marelli, (2021) were able to test phenomena posited by the morphological transcendence hypothesis (Libben, 2014). Second, CAOSS can operationalize semantic transparency compositionally. Here, Günther & Marelli, (2018) showed that measures *compositional* of semantic transparency (modifier and head composition), but not semantic relatedness, predicted lexical decision times (see also Günther et al., 2020; Günther & Marelli, 2020). Therefore, I propose that CAOSS can provide a quantitative characterization of the latent representational changes that are expected to emerge over time with prolonged exposure to compounds (Libben, 2014), and of the consequences these have on semantic transparency. Based on these considerations, the following hypotheses are advanced:

1. First, *semantic consistency* quantifies the similarity between the meaning of a constituent (e.g., "cat") and its role-bounded counterpart ("cat-"). Because the latter can be framed as the outcome of a discriminative process, greater consistency would reflect a state where the informative as-constituent meaning is closer to the original meaning, leading to greater

interference (or "semantic discrimination" efforts in Schmidtke et al., 2018) and increased processing times.

2. Second, *as-constituent relatedness* captures the degree of alignment between the to-be-accessed meaning of a compound and the predicted meaning that results from composition. Therefore, as-constituent relatedness should reflect the extent to which the compound meaning is preactivated as a result of compositional processes (analogous to "semantic boosting" in Schmidtke et al., 2018). On this account, as-constituent relatedness and processing times should be negatively correlated.
3. Third, the above measures both depend on as-constituents meanings, which are assumed to be learned via the repeated exposure to compounds. Thus, effects of semantic consistency and as-constituent relatedness should interact with measures of (item-level and individual-level) reading experience. Specifically, sensitivity to the former measures should increase with the latter.

Additionally, because compounding likely alters modifier and head meaning differently (figure 2.2; Günther & Marelli, 2021), one might observe some asymmetries between constituents (e.g., head constituents drift toward the compound more than modifiers, possibly leading to greater as-constituent relatedness effects; figure 2.2, right) Thus, the above hypotheses were tested separately for head and modifier constituents.

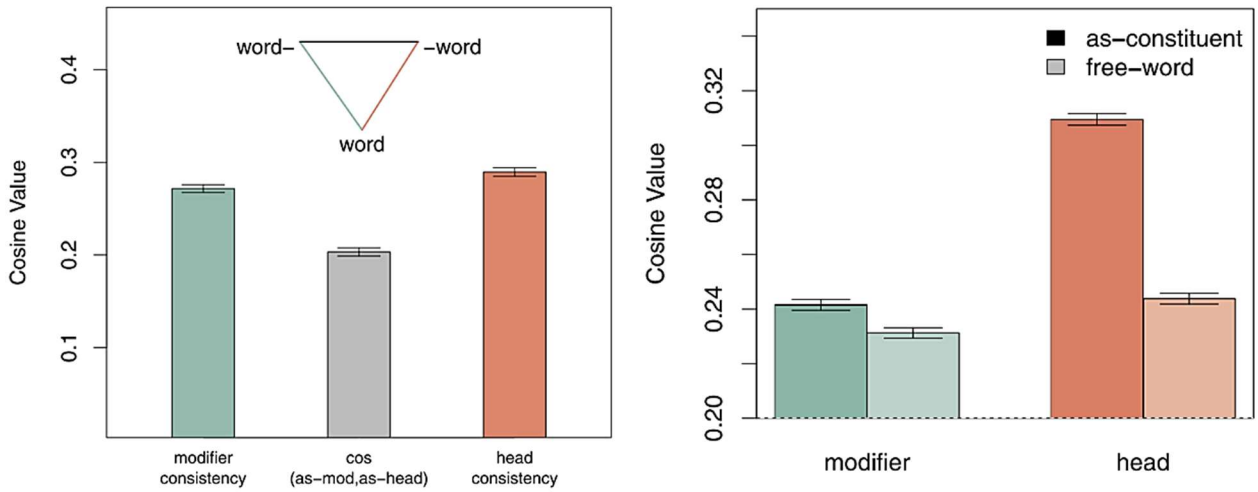

Figure 2.2: Modifier and head constituents differ in semantic consistency and as-constituent relatedness values (Günther & Marelli, 2021). *Left:* Modifier constituents are less consistent than heads. *Right:* As-constituent relatedness increases disproportionately more for head constituents.

To test these predictions, I framed this study in continuity with Schmidtke et al., (2018) and analyzed data from CompLex (Schmidtke et al., 2021), a large eye-tracking database of compound word reading for which measures of participants' reading experience are available.

2.2 Methods

2.2.1 Materials

I employed the CompLex dataset (Schmidtke et al., 2021), a combination of eye-tracking data from seven experiments devoted to compound word reading and encompassing eye movement variables from 440 participants, for which Author Recognition Test (ART; Acheson et al., 2008) scores (a measure of participants' exposure to printed materials) are also available. Across all the experiments that comprise the CompLex dataset, participants read a sentence containing one English noun-noun compound word; in some experiments, they were further tasked to answer comprehension questions (see Schmidtke et al., 2021 for additional description of the stimulus set and procedure). The full CompLex dataset encompasses reading data for 931 unique English noun-noun compounds, each unspaced and lexicalized (compounds were selected from the CELEX and ELP lexical database; Baayen et al., 1995; Balota et al., 2002). The database comprises eye-movement data for 47,763 trials. From this database, I excluded data for trials in which (a) ART scores were not available, (b) the compound word was skipped, and (c) compound or constituent words did not meet the criteria for CAOSS application (see the following section). Fixation times in the top 1% had been already excluded (Schmidtke et al., 2021). The resulting dataset consisted of 42,172 trials.

2.2.2 Modeling

Semantic Space definition: The free-word vectors of constituent and compounds were obtained from one of the semantic spaces provided by Baroni, Dinu, et al., (2014), where word vectors were derived from a ~ 2.8 billion word corpus obtained from the concatenation of the web-collected ukWaC (Baroni et al., 2009), a 2008 English Wikipedia dump, and the print-media-based British National Corpus (BNC consortium, 2007). Word vectors were derived with the *cbow* version of the *word2vec* model (Mikolov et al., 2013), a prediction-based DSM (Mandera et al., 2017) aimed at predicting a word given its context. The model was trained choosing the parameter set producing the best empirical results in Baroni et al., (2014), namely where 400-dimensional vectors were trained to predict words within a context window of five words (negative sampling with $k = 10$, subsampling with $t = 1e{-}5$). These word vectors represent the whole-word, lexicalized meaning of compounds, the free-word, lexicalized meaning of constituents, and form the basis on which CAOSS is trained.

CAOSS training: CAOSS was trained to predict a compound's whole-word vector from its constituents' free-word vectors. Thus, the training data consisted of a set of ordered triplets of (modifier, head, compound) vectors. Words forming the triplets belonged to a set of 4,429 unspaced compounds and their respective constituents. This set is the union of the English compound word databases of Gagné et al., (2019), Günther & Marelli, (2018), Juhasz et al., (2014) and Kim et al., (2019). Only unspaced compounds were considered, as CompLex only contains unspaced compounds (in line with previous studies employing CAOSS; Günther & Marelli, 2018, 2020, 2021; Marelli et al., 2017). From this set of compounds, only those for which the compound and its constituents had a frequency higher than 50 in the Baroni et al., (2014) corpus were retained in order to obtain reliable representations for all words. CAOSS parameters (i.e., the M and H matrices) were estimated with the DISSECT toolkit (Dinu et al., 2013a) by computing the closed-form solution of the least-squares linear regression problem (Dinu et al., 2013b).

2.2.3 Analyses

Dependent variables: Among the eye movement variables reported in CompLex (Schmidtke et al., 2021), I focused on first fixation duration (FFD; the duration of the first fixation on the compound), gaze duration (GD; the sum of the durations of all first-pass fixations on the compound) and total fixation time (TFT; the overall duration of all fixations on the word). These eye-movement variables can be related to the time course of word processing (Baayen et al., 2010; Kuperman et al., 2009). Specifically, FFD reflects the early processing of compound words when the first pieces of information become available; GD indicates the time required for the entire word-comprehension process, and TFT further accounts for processing difficulties associated with meaning integration (Marelli & Luzzatti, 2012). All fixation durations were logarithmically transformed to correct for the skewness of the distribution (Schmidtke et al., 2021).

Semantic predictors: I computed four semantic predictors from the set of whole-word and compositional representations defined by CAOSS (figure 2.1). Specifically, based on the hypotheses outlined in the introduction, the following measures were defined:

- *As-modifier/as-head relatedness:* The cosine similarity between a whole-word compound and the as-constituent meaning of one of its constituents. This measure represents the proximity of CAOSS predictions to a compound's actual (lexicalized) meaning. For example, "strawman" and "snowman" retain little of the meaning of "man", which simply indicates "something shaped like a man". If CAOSS generates a morphologically-bound representation

of “man” that conforms to this meaning, such unit, “-man”, should become closer in meaning to “strawman” and “snowman”, i.e., it should lead to greater as-head relatedness (but to lower as-head relatedness in the case of “policeman” and “fireman”, where “-man” should retain its original, free-word meaning). Greater values of as-modifier and as-head relatedness were expected to facilitate compound word processing resulting in shorter fixation times.

- *Modifier/head consistency:* The cosine similarity between a free-word constituent and its as-constituent meaning. Higher consistency indicates that the meaning of a word is similar to the one it takes on when it is part of a compound (e.g., “straw” and “straw-” in “strawman”); lower consistency reflects a greater change in meaning (e.g., “man” and “-man” in “strawman”). Greater consistency values were hypothesized to reflect greater interference between constituent and as-constituent meanings, resulting in longer fixation times.
- *Modifier/head relatedness:* The cosine similarity between a whole-word compound and one of its constituents. Following the above examples, the free-word meaning of “man” should be less similar to “strawman” (low head relatedness) than to “policeman” (high head relatedness). Since relatedness measures reflect the proximity of lexicalized, static representations, these measures were expected *not* to interact with measures of language experience. Moreover, following previous results (Günther & Marelli, 2018, 2020; see section 1.3.2), CAOSS-based measures were expected to provide superior operationalizations of semantic transparency effects, resulting in greater model fit.
- *Modifier/head composition:* The cosine similarity between a compositional compound and the free-word meaning of one of its constituents. This measure was included in continuity with previous studies, where it successfully predicted response times to lexical decisions (Günther et al., 2020; Günther & Marelli, 2018, 2020) and timed sensibility judgements (Günther & Marelli, 2020). These tasks require fast judgements of words presented in isolation, for which the semantic information required to perform the task is limited. Instead, participants in CompLex were asked to read and comprehend sentences that contained lexicalized compounds. Thus, accessing the actual whole-word compound meaning should be vital to perform the task properly. Since modifier and head composition do not consider actual whole-word (i.e., lexicalized) compound meanings, but as-constituent relatedness does, I expected the latter to better explain the effects of transparency in the context of CompLex.

Schmidtke, Van Dyke, et al., (2018) collapsed semantic discrimination and boosting onto the same dimension, namely the similarity between a compound and its constituents, relying on the effect of reading experience to tease the two apart. Instead, here discrimination and boosting were

operationalized by different constructs from the CAOSS framework (constituent consistency and as-constituent relatedness; figure 2.1), and reading experience is expected to affect both dimensions via the as-constituent representations they share. Moreover, they could also be considered statistically disentangled: the correlation between modifier consistency and as-modifier relatedness was moderate ($r = .33, t(866) = 10.219, p < .001$), as it was for head consistency and as-head relatedness ($r = .33, t(866) = 10.391, p < .001$).

Baseline predictors: In addition to the semantic measures described above, multiple predictors were employed as a non-semantic baseline[2]. The specific values for each predictor were obtained from the CompLex database (Schmidtke et al., 2021). Word frequency values reported the frequencies of occurrence of modifiers, heads, and compounds as stand-alone words in the SUBTLEX-US corpus (Brysbaert & New, 2009). Modifier and head family size measures were also included, which report the number of English compound words sharing either the left or right constituent, respectively. A Laplace transformation (i.e., increasing the raw frequency by 1 and taking the logarithm) was applied to the frequency and family size values (Brysbaert & Diependaele, 2013). To control for constituent differences that can be attributed to frequency, I further included, for each compound, the ratio of the log-transformed word frequency of its modifier to the log-transformed word frequency of its head (frequency ratio). I further included the orthographic length (in characters) of compound words, the plausibility of each compound word in its sentential context (plausibility ratings were collected with norming studies described in Schmidtke et al., 2018), and participants' ART scores (Acheson et al., 2008). As the raw ART scores were highly rightly skewed, a square root transformation was applied following Schmidtke et al.' (2018) procedure (figure 2.3). Pairwise correlations between all predictors are shown in figure 2.4.

Among these predictors, I considered compound word frequency and Author Recognition Test (ART; Acheson et al., 2008) scores as measures of reading experience at the item- and individual-level, respectively. Compound word frequency was takne to capture the amount of combinatorial knowledge that participants are expected to have with each item. While the effects of compound and constituent word frequency have been interpreted, respectively, as suggestive of whole-word and

[2] I attempted to frame this work in continuity with Schmidtke et al., (2021) and Schmidtke, Van Dyke, et al., (2018), employing the same baseline predictors of Schmidtke, Van Dyke et al., except for the inclusion of frequency ratio and the exclusion of predictability ratings. CompLex reports Cloze Probability values as measures of compound word predictability which, however, were equal to zero for every compound, as the stimuli were designed to render compounds unpredictable from their context Schmidtke et al., (2021). The predictability ratings of Schmidtke et al. were similarly mostly close to zero, indicating that compounds were unpredictable from their context.

decompositional processes, Baayen et al., (2007) argue that paradigmatic and syntagmatic cues narrow the set of possible lexical interpretations down to those that are morphologically valid (e.g., distinguishing between “hat” in “hatless” and “hat” in “chat”). Under this interpretation, complex word frequencies operate on this narrowed-down set of morphological family members (“hat+”): “Surface frequency reflects the combinatorial knowledge based on previous experience that is crucial for accepting the constituents as genuine meaning-bearing units rather than as strings providing accidentally partial matches.” (Baayen et al., 2007; p. 439). In this sense, the frequency of a compound word can be understood as the joint probability of its modifier and head constituent (Baayen et al., 2010).

On the other hand, ART scores were employed to measure the amount of experience each individual is expected to have with printed materials (Acheson et al., 2008; Moore & Gordon, 2014). The ART comprises a list of 130 names, half of which correspond to real published authors, which participants have to identify. Each participant’s score is computed as the total number of correct responses minus the number of errors; therefore, scores range from –65 to 65. ART scores correlate moderately with other measures of reading experience, strongly reflect the frequency of author names in text corpora, and predict gaze durations during sentence reading (Moore & Gordon, 2014). With specific reference to compound word reading, ART scores have been shown to predict various measures of eye movements as main effects (Falkauskas & Kuperman, 2015), as well as in interaction with other lexical and semantic predictors, most notably head transparency (Schmidtke, Van Dyke, et al., 2018) and compound word frequency (Falkauskas & Kuperman, 2015)[3].

[3] As a measure of participants' print exposure, Schmidtke, Van Dyke, et al., (2018) used the sum of participants’ scores on the ART and scores on the Magazine Recognition Test (Acheson et al., 2008). This composite measure of print exposure correlated .93 with ART. Moreover, Falkauskas & Kuperman (2015) retained ART scores as the only reliable measure of reading experience effects on eye movement data. Therefore, although I was limited in using ART scores as they were the only measure of reading experience available in the CompLex dataset (Schmidtke et al., 2021), this restriction should not be considered to be a significant limitation.

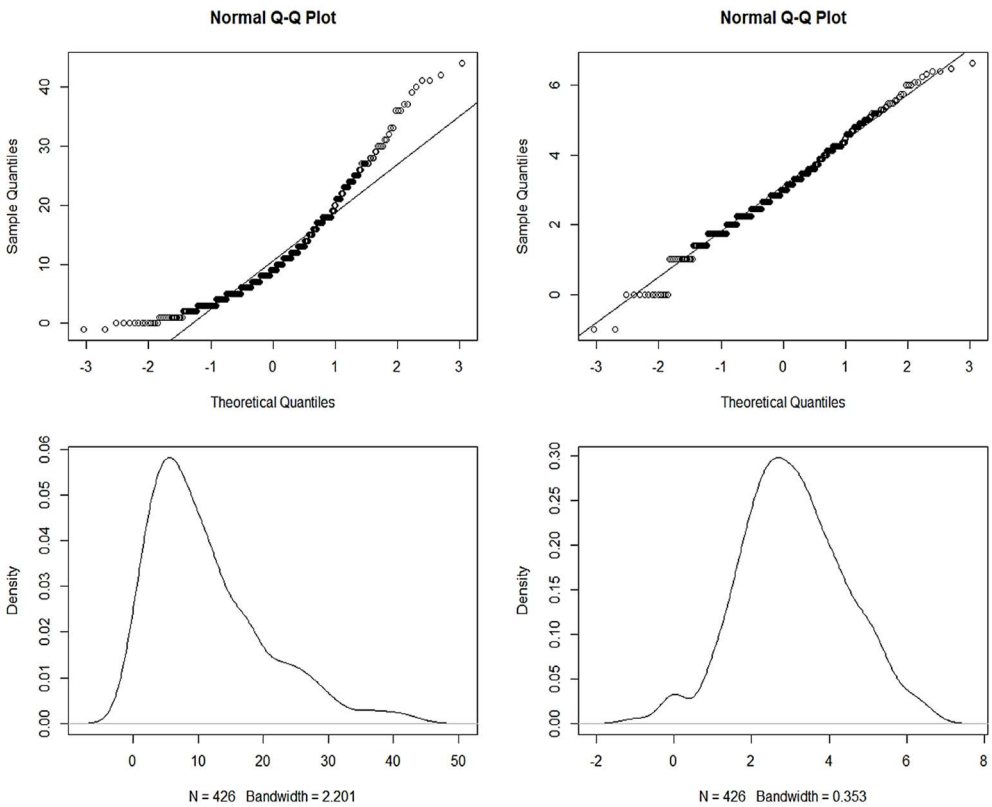

Figure 2.3: Author Recognition Test (ART) scores transformation. *Top left:* Q-Q plot of raw ART values. *Bottom left:* Density plot of raw ART values. *Top right:* Q-Q plot of ART values after square root transformation. *Bottom Right:* Density plot of ART values after square root transformation.

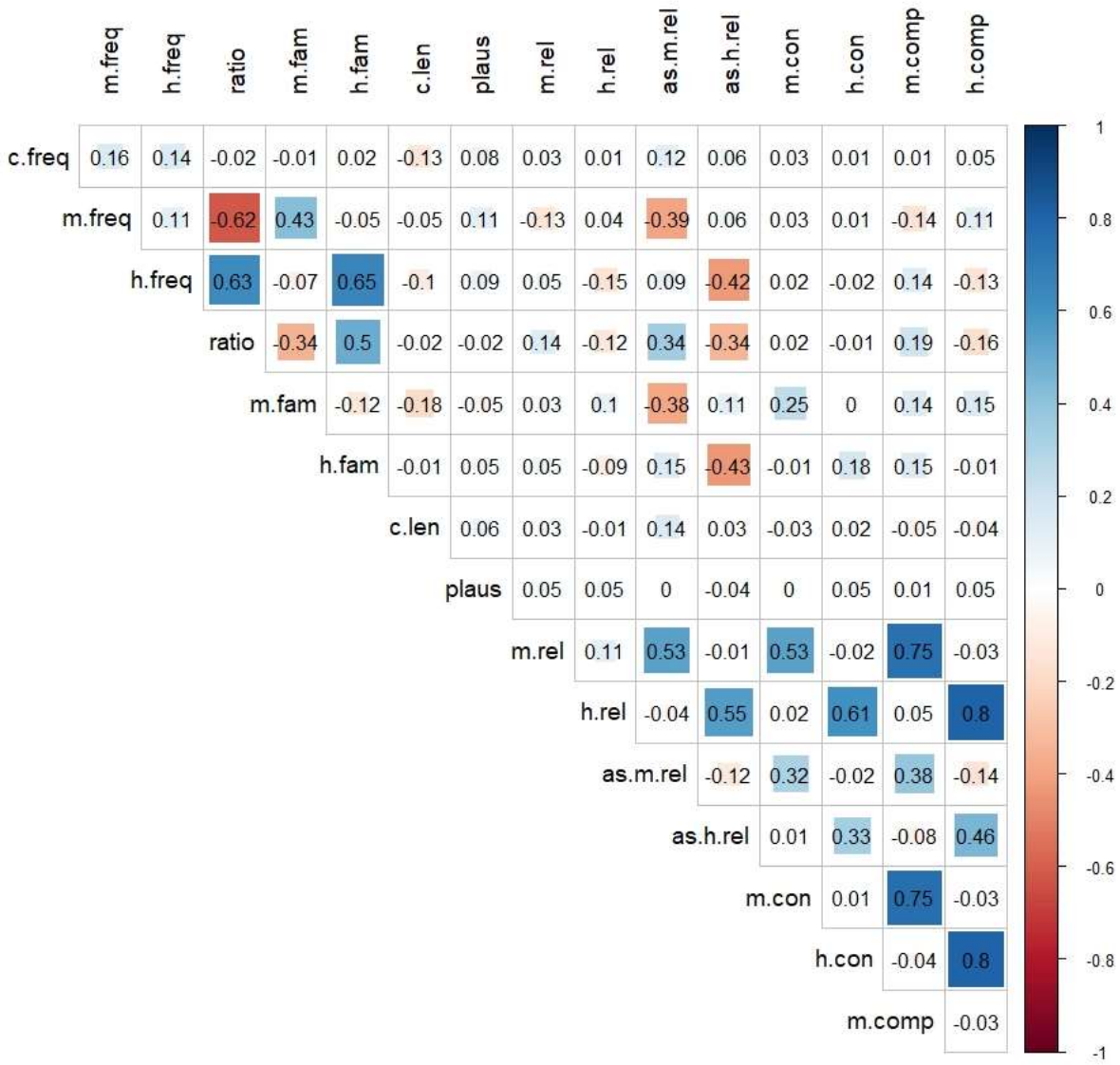

Figure 2.4: Pairwise correlations between measures of relatedness (rel), consistency (con), composition (comp), frequency (freq) and frequency ratio (ratio), family sizes (fam), and plausibility. Prefixes refer to compounds (c.), their modifiers (m.), and their heads (h.). Colored squares correspond to significant results (p <.05).

Statistical considerations: In each analysis, I started from a baseline linear mixed model containing all baseline predictors and random intercepts for modifiers, heads, compounds, and participant's identity. Then, a first set of analyses were conducted to test whether semantic predictors could further improve the baseline model and whether their effect was modulated by participants' reading experience. To do so, starting from the baseline model, each of the four semantic predictors was independently included in the statistical model in interaction with participants' ART scores. Likelihood-ratio tests were performed to assess whether the inclusion of a semantic predictor significantly improved the baseline. If more than one possible candidate predictor significantly improved its baseline model, the best-predicting model was identified based on its Akaike information

criterion (AIC) value (Akaike, 1998). This procedure was repeated until no further predictors significantly improved the model. This procedure allowed us to deal with the collinearity of the predictor variables (figure 2.4): additional variables can only improve model predictions when they explain additional variance, which cannot be due to the variance they share with other predictors already included in the model. Thus, when two predictors are highly correlated, their simultaneous inclusion in the model cannot be due to their overlap (Günther et al., 2020). I then performed a second set of analyses focusing on the influence of item-level reading experience, i.e., conducting the same set of analyses described above with compound word frequency in place of ART scores. Once the models were defined, I conducted model criticism to exclude the impact of overly influential outliers, removing residuals greater than |2.5| standard deviations and re-fitting on the truncated dataset (Baayen, 2008). Analyses and visualizations were performed with the packages lme4 (Bates et al., 2015), afex (Singmann et al., 2018) and corrplot (Wei et al., 2017) for the R software (R Core Team, 2017).

2.3 Results

First fixation duration: None of the semantic predictors improved the baseline model predicting FFD data. The addition of as-modifier relatedness in interaction with compound word frequency did not significantly improve the baseline model ($\chi^2(2, N = 42172) = 5.33, p = .070$).

Gaze duration: The addition of modifier consistency in interaction with participants ART scores significantly improved the baseline model describing gaze durations, $\chi^2(2, N = 42172) = 9.13, p = .010$. In this model, a significant interaction emerged between modifier consistency and participants ART scores ($b = .005, t(41535) = 2.462, p = .014$). Table 2.1 displays the parameters of the final model after model criticism.

When item-level reading experience was considered, the baseline model describing gaze durations was significantly improved by the addition of as-modifier relatedness in interaction with compound word frequency, $\chi^2(2, N = 42172) = 6.20, p = .045$. First non-significant ($b = -.004, t(708) = -1.872, p = .062$), the interaction between as-modifier relatedness and compound word frequency reached significance after model criticism. Table 2.2 displays the parameters of the final model after model criticism.

	Estimate	*t*	*p*

Baseline predictors			
Intercept	5.9938	117.8108	<.0001***
Compound log-frequency	-0.0455	-17.691	<.0001***
Modifier log-frequency	-0.0036	-0.5835	0.5598
Head log-frequency	-0.0141	-2.7991	0.0053**
Modifier family size	-0.0048	-0.8624	0.3891
Head family size	-8e-04	-0.1342	0.8934
Compound length	0.0446	10.1598	<.0001***
Plausibility	-0.0065	-1.6185	0.1059
ART	-0.0957	-10.058	<.0001***
Frequency ratio	0.0362	1.2186	0.2235
Semantic predictors			
Modifier consistency	*-0.0084*	*-1.7993*	*0.073*
ART × Modifier consistency	**0.0062**	**3.1853**	**0.0014****

Table 2.1: Parameters of the final model best-describing gaze duration when semantic predictors are allowed to interact with participants Author Recognition Test scores. The table displays parameters after model criticism.

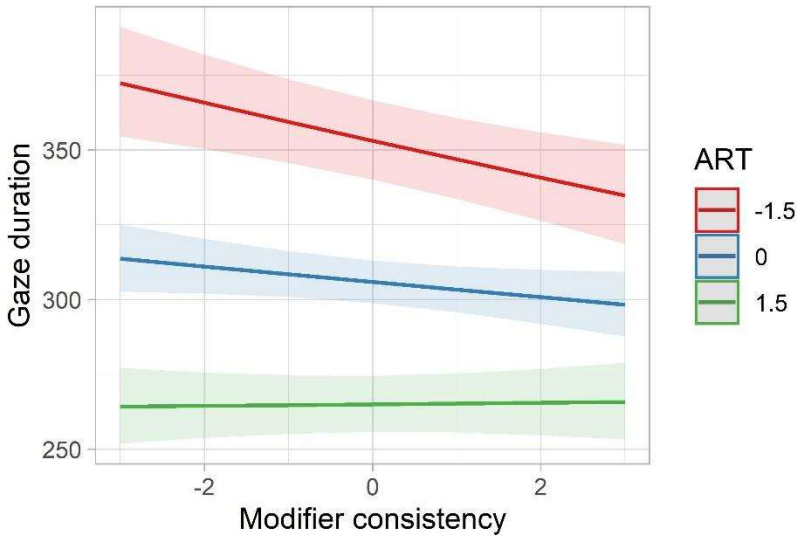

Figure 2.5: Interaction effects between modifier consistency and participants ART scores on gaze durations.

	Estimate	*t*	*p*

Baseline predictors			
Intercept	6.0188	118.3613	<.0001***
Compound log-frequency	-0.0447	-17.0559	<.0001***
Modifier log-frequency	-0.006	-0.9783	0.3284
Head log-frequency	-0.0128	-2.5436	0.0112*
Modifier family size	-0.0092	-1.685	0.0931
Head family size	-3e-04	-0.0499	0.9603
Compound length	0.0453	10.2758	<.0001***
Plausibility	-0.0071	-1.7738	0.0764
ART	-0.0954	-10.0202	<.0001***
Frequency ratio	0.0267	0.9025	0.3672
Semantic predictors			
As-modifier relatedness	0.0088	0.971	0.3319
Compound log-frequency × As-modifier relatedness	**-0.0048**	**-2.0474**	**0.041***

Table 2.2: Parameters of the final model best-describing gaze duration when semantic predictors are allowed to interact with compound word frequency. The table displays parameters after model criticism.

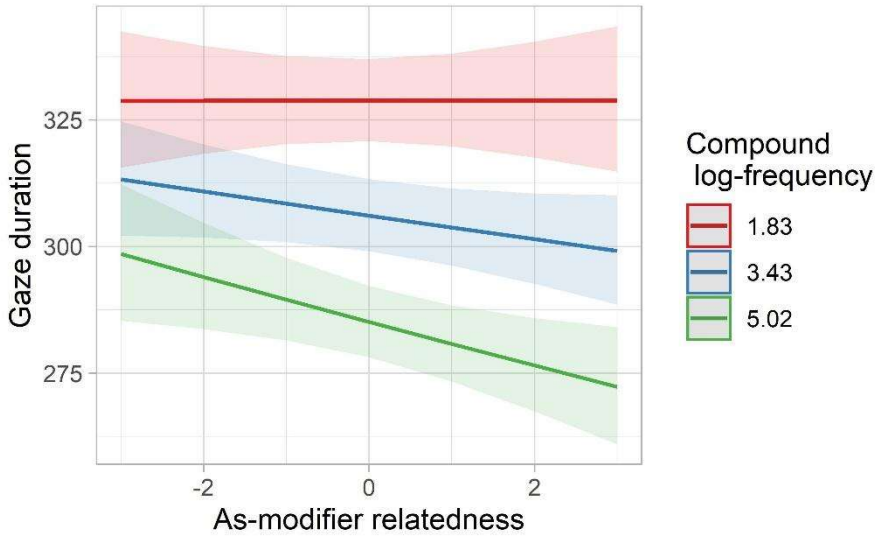

Figure 2.6: Interaction effects between as-modifier relatedness and compound word frequency on gaze durations.

Total fixation time: The baseline model predicting TFT failed to converge. The all_fit() function from the afex package (Singmann et al., 2020) was used to re-fit the baseline model with multiple optimizers (Brown, 2021). No convergence warnings or singular fits were observed across all optimizers. Therefore, the linear mixed model optimizer was changed to *bobyqa*, and the analyses were re-run with this optimizer. None of the semantic predictors improved the baseline model predicting TFT data in interaction with compound word frequency. Instead, a significant improvement over the baseline model was observed after the inclusion of as-modifier relatedness in interaction with participants ART scores, ($\chi^2(2, N = 42172) = 10.35, p = .006$). A significant main effect of as-modifier relatedness was observed (figure 2.7; $b = -.020, t(699) = -3.077, p = .002$), but the interaction of as-modifier relatedness and participants ART scores was not significant ($b = -.022, t(41361) = -0.959, p = .337$). Table 2.3 displays the parameters of the final model after model criticism.

	Estimate	*t*	*p*
Baseline predictors			
Intercept	6.3125	89.2982	<.0001***
Compound log-frequency	-0.0647	-17.0627	<.0001***
Modifier log-frequency	0.0021	0.2401	0.8104
Head log-frequency	-0.017	-2.3692	0.0181*
Modifier family size	-0.0089	-1.2	0.2311
Head family size	0.0113	1.4113	0.1597
Compound length	0.0543	8.7808	<.0001***
Plausibility	-0.0269	-4.8117	<.0001***
ART	-0.0972	-8.5457	<.0001***
Frequency ratio	0.0549	1.2945	0.196
Semantic predictors			
As-modifier relatedness	**-0.0217**	**-3.2772**	**0.0011****
ART × As-modifier relatedness	-0.0019	-0.8812	0.3782

Table 2.3: Parameters of the final model best describing total fixation time when semantic predictors are allowed to interact with participants Author Recognition Test scores. The table displays parameters after model criticism.

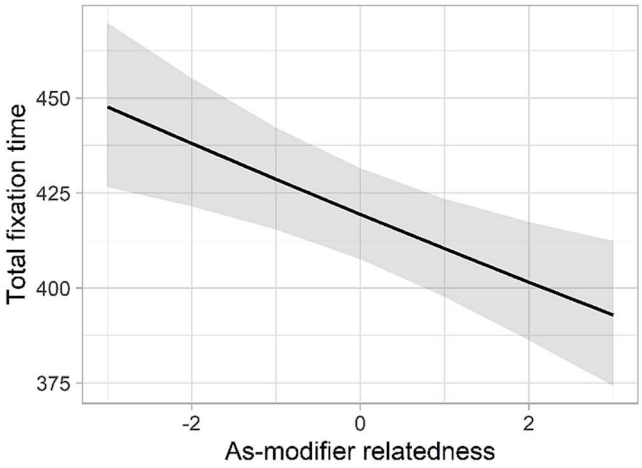

Figure 2.7: Main effect of as-modifier relatedness on total fixation time.

2.4 Discussion

Leveraging the CAOSS framework, latent, as-constituent semantic representations were generated and used to operationalize and disentangle inhibitory and facilitatory effects of semantic transparency and consistency (i.e., semantic boost and discrimination; Schmidtke, Van Dyke, et al., 2018). Under the assumption that as-constituent representations emerge with prolonged exposure to compound words (Libben, 2014), I tested the interactions among measures of transparency and reading experience.

As-constituent relatedness: The effects of as-constituent relatedness conforms to hypotheses 2 and 3. First, I observed a facilitatory main effect of as-modifier relatedness on TFT: greater values of as-modifier relatedness were associated with shorter fixations. Because CAOSS re-weights the features of a constituent according to their informativity for compounding, as-modifier relatedness can be considered as an index of compound meaning predictability given the modifier. Thus, the facilitation observed is interpreted as evidence for the processing benefits arising from greater predictability. Second, I observed an interaction effect of as-modifier relatedness and compound word frequency on GD. This finding complements the first by showing that the facilitation stemming from predictability increases with greater compound word frequency. According to Baayen et al., (2007), the frequency of a compound word reflects the amount of combinatorial knowledge available to identify the constituents as meaning-bearing units (or the joint probability of the modifier and head combination; Baayen et al., 2010). Therefore, this interaction can be taken to indicate that the greater

the evidence to process the compound compositionally is, the more one can experience the processing benefits that stem from successful composition.

Semantic consistency: First, main effects of semantic composition were absent on any eye movement variable. However, an interaction emerged between modifier consistency and participants ART scores on GD: the effect of modifier consistency, initially facilitatory, weakens with the increased ART scores. Although the shape of the interaction is in line with the hypothesis (i.e., increasing inhibition from semantic consistency with increasing reading experience), it is observed, contrary to my predictions, that participants with *lower* ART scores are more sensitive to modifier consistency and, in addition, are initially *facilitated* by higher consistency. High semantic consistency indicates that free-word constituent meaning is already predictive of the meaning it will take on within a compound. For example, "news" typically retains its meaning when used as a modifier (e.g., "newspaper", "newsletter"; the modifier consistency of "news-" is 0.53), while "cat" does not (e.g., "catfish", "catwalk"; the modifier consistency of "cat-" is 0.16). Therefore, the observed interaction could be interpreted as follows. If morphologically transcended (as-constituent) representations are learned over time with prolonged reading experience (Libben, 2014), participants with lower reading experience would rely less on them and more on the available free-word meanings. While high consistency ("news" and "news-") suggests that free-word meanings are already informative for compounding – and are therefore incidentally available to less proficient readers – only more proficient readers would be equipped to deal with low consistency cases ("cat" and "cat-"), where the as-constituent meaning must be correctly inferred. This could explain why readers differ more at the low end of the consistency spectrum. On this account, participants with high reading experience would have learnt to rely on the transcended (as-constituent) meanings, irrespective of their proximity to the free-word, and would thus be insensitive to this dimension.

Operationalizing semantic transparency: Consistent with my predictions, across all models, compositional measures outperformed semantic relatedness, replicating the finding that semantic transparency is best operationalized compositionally (Günther & Marelli, 2018). Moreover,, because participants needed to access whole-word compound meaning in order to understand the sentences in CompLex, I predicted that as-constituent relatedness would be more apt at capturing reading behavior compared to modifier or head composition, which focus on compositional compound meaning that might not be interpretable, but sufficient to influence shallower lexical decision tasks (as tested in previous studies). Indeed, as-constituent relatedness consistently outperformed composition measures. This might explain why, in previous studies, modifier and head composition have been able to predict response times for lexical decision and sensibility judgment tasks (Günther et al., 2020;

Günther & Marelli, 2018, 2020) but not eye movements (Günther et al., 2020). This result shows that different operationalizations of semantic transparency can be more or less adequate depending on the experimental setting.

CAOSS and morphological transcendence: In line with previous studies (Falkauskas & Kuperman, 2015; Schmidtke et al., 2018; Schmidtke, Van Dyke, et al., 2018), these findings can be seen as stemming from constituent binding, namely the increasing intertwining of a word's orthographic, phonological, and semantic characteristics posited by the Lexical Quality Hypothesis (Perfetti, 2007; Perfetti & Hart, 2002). Analogously, the morphological transcendence hypothesis (Libben, 2014) posits the progressive development of constituent-specific word meanings with increasing exposure to these lexical items. Both constructs can be formalized by as-constituent representations within the CAOSS framework (which is, however, limited to the semantic and lexical domains). On the one hand, when the goal of compound processing is the retrieval of its whole-word meaning, as-constituent relatedness would capture the degree of facilitation granted by as-constituent representations (e.g., exploiting the meaning of "cat-" to retrieve the meaning of "catfish"). On the other hand, however, contrary to the original hypothesis, the present results call for a different conceptualization of the development of morphologically transcended units. Initially, I hypothesized that prolonged exposure to a constituent would lead to the development of its as-constituent representation which, while informative of compound word meaning, would interfere with its free-word counterpart, assuming both meanings would be simultaneously available. Instead, the present results are more compatible with the idea that learners gradually shift from free-word to as-constituent meaning. On this view, the two are not available simultaneously, and as-constituent meanings are informative representations that readers learn to access or generate. These considerations are aligned with a dynamic, context-dependent view of lexical semantics, where morphological information constitutes one form of constraint on the process of settling into the appropriate (compound) interpretation (Libben, 2017). Therefore, whether one considers as-constituent representations to be the result of a mechanism of automatic combination, improved lexical quality, or morphological transcendence, results show that they constitute an informative unit of analysis, not only with respect to compound representations (Günther & Marelli, 2021) but also concerning their processing.

In a previous study, Günther & Marelli, (2021a) demonstrated that the CAOSS framework is well suited to model morphologically transcended units (Libben, 2014). Indeed, by modeling compounding as a predictive process mediated by as-constituent representations, the latter can be naturally framed as morphologically transcended units. In this chapter, this line of research is further extended in two ways. First, whilst Günther & Marelli, (2021a) focused on explicit judgements

(ratings), this study focused on chronometric data, testing whether as-constituents affect compound processing beside representation. Second, this study tested an additional tenet of the morphological transcendence hypothesis, namely that role-bounded representations are developed with prolonged (compound) reading experience. Results show that as-constituents do affect processing, confirming the superiority of compositional measures of transparency, and their effect interacts with item- and individual-level measures of experience, consistent with the idea that combinatorial knowledge evolves over repeated exposure to compounding.

Chapter 3

Predicting conceptual combination by mapping language embeddings to visual representations

3.1 Introduction

According to theories of embodied cognition, sensorimotor experience is not only fundamental for our interaction with the world and for the way we learn conceptual representations; rather, the same information is crucial for *representing* conceptual information (Barsalou, 2008, 2020). Consequently, on this account, framing conceptual combination from a purely linguistic or "amodal" perspective is limiting (Wu & Barsalou, 2009). Indeed, according to the Embodied Conceptual Combination (ECCo) model (Lynott & Connell, 2010), conceptual combination relies on two streams of information and cognitive processes, linguistic distributional information and situated (sensorimotor) simulations, which are hypothesized to interact dynamically through time to determine the appropriate combination of constituent concepts, with their relative importance dependent on the specifics of the stimulus, its context and the task at hand. According to the ECCo model, concepts can be combined *destructively* (i.e., where the meaning of the modifier, but not the head, is reduced to one or few specific properties; e.g., "swordfish" as "a fish with a bill shaped like a sword", where the concept of "sword" is reduced to its shape) or *non-destructively* (i.e., where the meaning of both modifier and head constituents is largely preserved; e.g., "swordsman" as "a man who fights with a sword"), with some combinations being amenable to both interpretations (e.g., "cactus beetle" as "a beetle with spikes" or "a beetle that eats cacti"). Formal accounts of how the language and sensorimotor systems could interact in service of conceptual combination are, however, still lacking.

Moving from general theories of cognition to computational cognitive models, the problem of grounding has been central to distributional semantics models (Günther et al., 2019), and to their deep learning descendants (Bender & Koller, 2020). While some studies show that distributional models capture extralinguistic properties akin to world models (e.g., Gatti et al., 2022; Li et al., 2022), other efforts have been devoted to the definition of semantic spaces grounded in extralinguistic

information. Pursuing the latter, Günther et al., (2022) define *vision spaces* based on the representations of target layers of VGG-F (Chatfield et al., 2014), a convolutional neural network (CNN) for image classification. CNNs are deep learning models trained to predict the correct label of an image by encoding its features in embedding spaces, which are transformed through a series of layers culminating in a classification layer, namely a list of categories to which the image is assigned (see general introduction, section 1.1.4). In a large-scale study, Günther et al., (2022) demonstrate that the similarity between vision embeddings predicts human ratings of image similarity, visual word similarity and image typicality, as well as chronometric measures. Thus, vision spaces of this kind open up the possibility of defining quantitative models of visually-grounded conceptual combination. Indeed, Günther et al., (2020) trained a CAOSS model on vision embeddings, demonstrating that vision-based transparency measures predict compound word processing over and above language-based ones. However, besides building combinatorial models that operate directly on vision features, another possibility is to combine constituents in a language space, and then infer visual features thanks to the structural similarities that link the same concepts in different modalities (Roads & Love, 2020). Indeed, multiple studies have demonstrated that the embedding spaces of vision and language models, while trained independently on different modalities and objectives, can be related with a simple linear mapping function (Günther, Petilli, Vergallito, et al., 2020; Lazaridou et al., 2014; Li et al., 2023; Merullo et al., 2022; Sorscher et al., 2022). This type of cross-modal alignment appears to support zero-shot generalization (Aho et al., 2022; Sorscher et al., 2022), namely inferring the properties of a stimulus in a new domain by exploiting what is known of the stimulus in another (e.g., learning visual concepts by relying on linguistic descriptions and symbols; Lowerse, 2018). For example, Günther, Petilli, Vergallito, et al., (2020) exploited these structural similarities to linearly map the embeddings of abstract words from a language space onto a vision space, showing that the predicted vision embeddings possessed interpretable semantic neighbors (images). This raises the possibility that a similar strategy might be leveraged to infer the perceptual properties of novel combinations. Specifically, the idea that motivates the present Chapter is that the visual features of novel combined concepts can be predicted through the sequential application of language composition and language-to-vision mapping. In particular, this Chapter advances three hypotheses about the type of contribution that language might exert on (visual) conceptual combination:

1. *Language captures the interpretation of novel (visual) compounds*: Distributional language information predicts which of multiple interpretations of the same novel compound is considered more plausible;
2. *Language captures the typicality of novel (visual) compounds:* Language does not simply capture which interpretation is preferred, but also which image *of the same* interpretation

is preferred. Under this stronger hypothesis, language not only captures conceptual, but also markedly visual information;

3. *Language predicts the processing times in response to novel (visual) combinations:* While the above hypotheses concern untimed, explicit choices, it is possible that language predictions only influence processing, and its effects are thus detectable only when looking at chronometric behavior.

To test these hypotheses, I first build a dataset of 80 novel noun-noun combinations (e.g., "doughnut pizza"), each receiving a destructive interpretation ("a pizza with a hole in the middle") and a non-destructive interpretation ("a pizza with doughnut topping"). For each interpretation of each combination I then use a text-to-image model (Betker et al., 2023) to generate corresponding synthetic images. In an online experiment, participants are presented with these novel combinations as written words. They are asked to anticipate their visual appearance and to later select, in a forced choice task, which of two images best matches their intuition.

The CAOSS model is then used to model language combination, while CNNs are used to represent visual information (Günther et al., 2023). Then, compound embeddings are linearly mapped from the language space to the vision space, and the mapped embeddings are taken to be their predicted visual representation (Figure 3.1). Then, the similarity between each image and the predicted embedding is computed. Following the three hypotheses outline above, it is expected that the similarities between an image and the predicted embedding will be such that (1) the image depicting the preferred interpretation will be closer to the predicted embedding, (2) the chosen image *of the same interpretation* will be closer to the predicted embedding, (3) the time taken to select the correct image will be negatively proportional to its proximity to the predicted embedding.

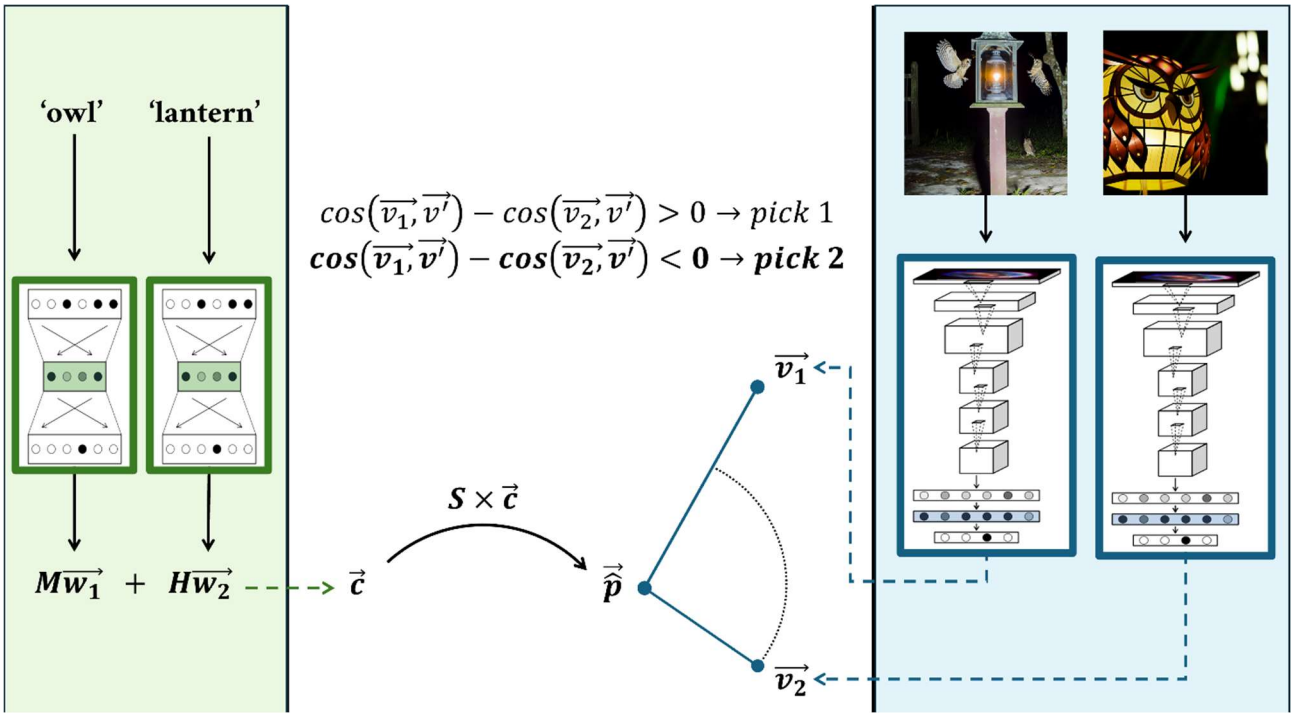

Figure 3.1: Overview of the present study. The constituents of a novel compound like "owl lantern" are embedded in the word2vec language space and combined with the CAOSS model. The resulting embedding, $\vec{c}$, is linearly mapped by the S function to the vision space defined by a target layer of a convolutional neural network. The so mapped embedding, $\hat{\vec{p}}$, is the language-predicted visual embedding of "owl lantern". Given two images depicting "owl lantern" (exemplars 1 and 2), these are fed to the CNN, and their embeddings, $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$, are extracted at the target layer. Then, the cosine similarity between $\overrightarrow{v_1}$ and $\hat{\vec{p}}$,, and between $\overrightarrow{v_2}$ and $\hat{\vec{p}}$, is computed. I predict that the exemplar that is closest to $\hat{\vec{p}}$, will be selected by human participants in a forced choice paradigm.

3.2 Methods

3.2.1 Models

Language space: Word embeddings are generated with the word2vec *cbow* model (Mikolov et al., 2013) using the parameter set configuration producing the best empirical results in Baroni et al., (2014): 400-dimensional vectors, a context window of five words with negative sampling with k = 10 and subsampling with t = 1e−5. The model was trained on a ~ 2.8 billion word corpus obtained from the concatenation of the web-collected ukWaC (Baroni et al., 2009), a 2008 English Wikipedia dump, and the print-media-based British National Corpus (BNC consortium, 2007).

Language composition: In order to obtain compound embeddings from those of constituent words, the CAOSS model is used (Marelli et al., 2017). The CAOSS model was trained on a set of (modifier, head, compound) embedding triplets from the following datasets of closed-form and

hyphenated compounds[4]: Balota et al., (2007), Gagné et al., (2019), Günther & Marelli, (2018), Juhasz et al., (2014), Kim et al., (2019), and Baayen, Piepenbrock, & Gulikers, (1995). To ensure embedding quality, only the triplets where all words appeared at least 50 times in the Baroni et al., (2014) corpus were considered. The training set consisted of 7352 triplets, and the model was trained with the DISSECT toolkit (Dinu et al., 2013).

Vision spaces: I employed vision spaces from the layers and CNNs showing the best results in modeling human rating data in Günther et al. (2023). These are GoogleNet (Szegedy, 2015) 5a_3x3_reduce layer (best prediction of image similarity and visual word similarity ratings), DarkNet-53 (Redmon & Farhadi, 2018) conv50 layer (best prediction of image typicality ratings) and VGG-M-1024 (Chatfield et al., 2014) fc7 layer (best mean prediction across image similarity, visual word similarity and image typicality ratings). All CNNs were pretrained on the ILSVRC2012 challenge datasets (the 1000 concept classes are consistent since the 2012 edition; Deng et al., 2009; Russakovsky et al., 2015). Note that a vision space is comprised not only of the embeddings of images from a given dataset (e.g., the training set); rather, new images can be fed to the CNN, and their embeddings extracted.

3.2.2 Synthetic images validation

The existing literature on the cognitive plausibility of vision spaces is based on evaluations conducted on real images. Thus, to ensure that the synthetic nature of the image stimuli does not lead to significant distortions in the representational space of CNNs, the following validation was conducted.

Real and synthetic images: First, a *validation set* was defined, namely a set of concepts for which real images are available, and for which synthetic images were generated for comparison (Figure 3.2). Because I was ultimately interested in synthetic images of novel compound words, this validation set should have answered the following question: "Assuming one had real images depicting novel combinations, would synthetic and real images for these novel combinations deviate in a given vision space?". Because real images for novel combinations are by definition unavailable, a validation should be devised in such a way that its results could be generalized to the case of novel combinations. To do so, I leveraged the fact that words that are absent from a CNNs training set can be considered

[4] When compound embeddings were available for both closed and hyphenated forms, we considered the most frequent variant in our corpus.

"novel concepts" for the network. Indeed, all CNNs were trained on the 1000 ImageNet-1k concept classes for the ILSVRC2012 challenge, a subset of the much larger ImageNet database (i.e., ImageNet-21k; Deng et al., 2009), which includes many *compounds* that do not appear in the CNNs' training set. Because one might expect some representational differences even among sets of real images coming from different datasets, ImageNet-21k was complemented with THINGSplus (Stoinski et al., 2023), a database of high-quality naturalistic images of 1854 concrete object concepts (concreteness rating ≥ 4; Brysbaert et al., 2014)[5]. After applying a set of criteria aimed at ensuring image and embedding quality (see Appendix A.1.1), the validation set comprised 424 concepts (212 compound words and 212 simple words). Then, OpenAI DALLE-3 (Betker et al., 2023), a large diffusion model (Zhang et al., 2023), was used to semi-automatically generate 9 synthetic images for each concept, from which 3 images deemed valid[6] were manually selected (Appendix A.1.1). The final validation set comprised 404 concepts (407 simple words and 197 compound words); each was represented 13 THINGSplus images, 13 ImageNet-21k images, and 3 synthetic images whose size always exceeded the largest CNN input size (i.e., $256 \times 256 \times 3$, DarkNet-53).

[5] Because the source images of both databases overlap, I ensured that no duplicate images could be present across datasets by computing the Euclidean distance between images in pixel space and manually inspecting the pairs with lowest distance. Only two duplicates were found. In this case, these images were kept in THINGSplus set and removed from the ImageNet set, ensuring that at least 13 ImageNet images were still available after removal.

[6] Images were judged by three raters (the first author and two collaborators). Images selected by more raters were preferred; in case of ties, preference was given to the first author's judgment. For an image to be selected, it had to depict at least partially, but always recognizably, the target concept in a realistic (i.e., photographic) style. When more than 3 images were deemed valid, we preferred images with complex, realistic backgrounds, as well as those that differed the most among each other, in order to increase their heterogeneity.

Figure 3.2: Example images from the validation set.

Comparing real and synthetic images: For the THINGSplus and ImageNet-21k stimuli, 10 images were randomly selected and averaged to define the prototype of each concept, and the remaining 3 constituted exemplar images. Thus, each concept was described by the following representations: a prototype embedding based on images from i) THINGSplus and ii) ImageNet-21k, and three exemplar embeddings based on iii) THINGSplus, iv) ImageNet-21k, and v) DALLE-3 images. Results are focused on the best CNN, DarkNet-53 (see section 3.2.3 for model selection). Real and synthetic images were compared in three conditions:

- *True-prototype ranking:* Exemplars should be closest to their true prototype, and synthetic image exemplars should not deviate substantially from this pattern. To test whether they do, the cosine similarity between the exemplars and all prototypes was computed, obtaining the ranking of the true prototype. Then, I computed the average rank of each true prototype, and compared the ranking based on real and synthetic exemplars. Using the THINGSplus prototype as a reference, the median rank of THINGSplus, ImageNet-21k and DALLE-3 exemplars were 1, 2 and 1, respectively (mean ranks were 2.0, 18.5, and 6.6). Instead, using ImageNet-21k prototypes, the median ranks were 1, 2 and 2, respectively (means were 9.4, 12.1, and 11.9).
- *Exemplar-prototype similarity:* Real image exemplars are closer to their prototype, and synthetic image exemplars should not substantially deviate from this pattern. Thus, I averaged the cosine similarity between all the exemplars and their respective prototype. Using the

THINGSplus prototype as a reference, the mean cosine similarity between the prototype and THINGSplus, ImageNet-21k and DALLE-3 exemplars were 0.45, 0.30 and 0.37, respectively. Using the ImageNet-21k, they were 0.35, 0.32 and 0.33.

- *Odd-one-out judgements:* The embeddings obtained from real images are known to predict behavioral measures, and embeddings obtained from synthetic images should be similarly predictive. Using a set of 4.70 million odd-one-out judgements, Hebart et al. estimated the similarity between all THINGS concepts. To test whether those align with CNNs' estimates of image similarity, I obtained the embedding of prototypes computed over 3 randomly chosen THINGS images, 3 randomly chosen ImageNet-21k images, and the 3 DALLE-3 images. In each case, I computed the cosine similarity between the prototypes of all concept pairs, populating a 404×404 similarity matrix. Then, I performed a representational similarity analysis (RSA; Kriegeskorte et al., 2008) between these similarity matrices and Hebart et al., (2023) similarity matrix, resulting in Spearman correlation coefficients of $\rho = 0.26$ (THINGSplus images), 0.26 (ImageNet-21k images) and 0.28 (DALLE-3 images).

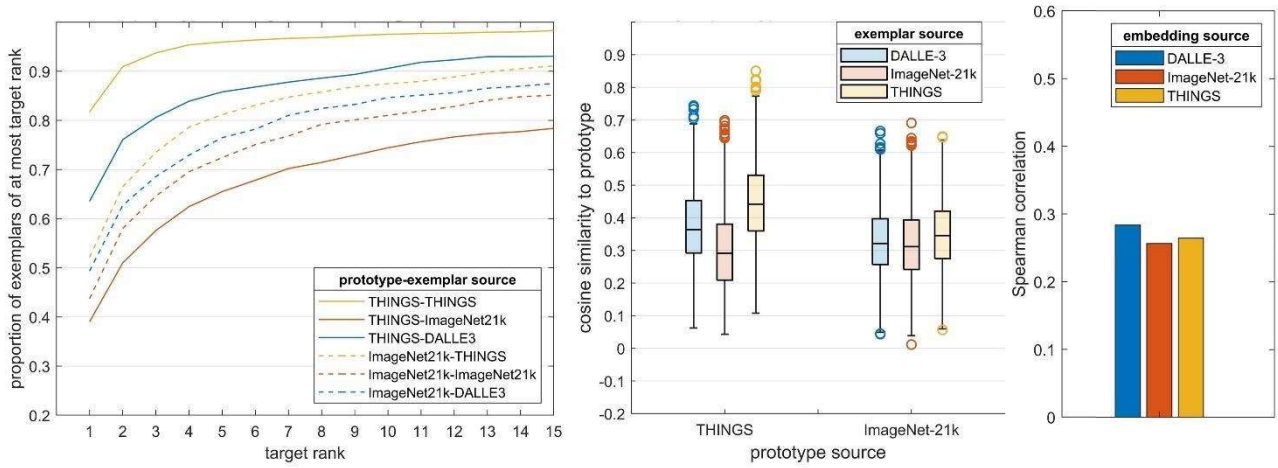

Figure 3.3: Synthetic image validation results for the DarkNet-53 model (at layer conv50) in a validation space of 404 concepts. (left) Proportion of exemplars for which the true prototype was among the *k* closest prototypes, plotted as a function of *k*. (middle) Boxplots of cosine similarities between prototypes and exemplars. (right) Spearman correlation coefficients between image similarity matrices generated with DarkNet-53 and the (estimated) behavioral similarity matrix from Hebart et al., (2023).

Overall, the same pattern of results is observed in the first two analyses, regardless of prototype source: exemplars based on THINGSplus images performed the best, followed by DALLE-3 exemplars and, finally, by ImageNet-21k exemplars (figure 3.3). In the third analysis, instead, DALLE-3 showed the highest second-order similarity with Hebart et al., (2023) similarity estimates. The underperformance of ImageNet-21k might be driven by the fact that, compared to THINGSplus

and DALLE-3 images, concepts are not always the most salient part of the image, and are not always depicted in canonical ways (e.g., unnatural perspective)[7]. Due to the lack of additional real image datasets meeting our criteria (section 3.2.2), I could not perform a large-scale comparison supporting quantitative analyses, and the above observations remain qualitative. Nonetheless, based on the consistent outperformance of ImageNet-21k and the similarity with THINGSplus (especially when the reference prototype is defined based on ImageNet-21k), I concluded that, for similarity-based comparisons, the embeddings of synthetic images generated by DALLE-3 are amenable to those based on real images. Having established the robustness of CNNs to synthetic images, we turn to connecting vision and language spaces.

3.2.3 Language-to-vision mapping

Multiple mapping functions were trained to align language representations to their counterparts in vision spaces. In order to assess the generalization of the mapping function to the domain of novel combinations *and* synthetic images, data was split into the following training, development and test sets:

i) **Training set**: This set comprised all concepts in the THINGSplus dataset which did not belong to the development and test sets. The same selection criteria used for the validation set (Appendix A.1.1) were applied, resulting in 1285 training concepts.

ii) **Development set**: The purpose of the development set was to select the hyperparameter values leading to accurate mapping functions. Because this function would ultimately generate embeddings to be compared to those of *synthetic* images (figure 3.1), I performed hyperparameter selection using the embeddings of synthetic images, and focused on the 207 simple word concepts from the synthetic images generated for the validation (section 3.2.2).

iii) **Test set**: The test set comprised the embeddings of the synthetic images of 196 compound words, again from the synthetic image used for validation. Note that the test set *did not* include images of the novel compounds used in the actual experiment. Language

[7] With this observation I do not suggest that ImageNet lacks validity or provides a weak comparison. Indeed, ImageNet is arguably the most popular reference for studies on naturalistic images, and previous large scale behavioral evaluations of CNN-based vision spaces relied on ImageNet (Günther et al., 2022, 2023).

embeddings were generated with CAOSS. Performance on the test set was used to select the CNN.

Based on evidence for the good performance of linear functions for language-vision mappings (Lazaridou et al., 2014; Li et al., 2023; Merullo et al., 2022; Sorscher et al., 2022), a linear regression model was trained. Moreover, because of the limited training set size and the high dimensionality of the embedding spaces, regularization based on the elastic net method was applied. Elastic net applies a linear combination of the L_1 and L_2 penalties (whose relative importance is controlled by the parameter α) with regularization coefficient λ. The α values employed were $\{0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$; for each, 20 λ coefficients were automatically selected by the *lasso* MATLAB function. The α and λ parameters were systematically varied during training, and those leading to the best-performing configuration on the development set were selected (figure A.1.1). Performance was evaluated based on the extent to which the predicted embeddings could be used in place of "true prototypes", namely the average embeddings of multiple images of the same concept. Specifically, we tested whether comparisons between the same pairs of DALLE-3 exemplar images were consistent across the two prototype types (figure 3.4). As discussed in the introduction, I was interested in predicting the type of interpretation of novel combinations (destructive vs. non-destructive, or *between-class* judgments) and the prototypicality of alternative depictions of the same interpretation (*within-class* judgments). To select a mapping optimized for those objectives, the following measures of mapping performance were computed (Figure 3.4):

a) *Between-class*: For a given concept class j, its prototype p_j was defined by averaging the embeddings of 10 THINGSplus images depicting that concept. Then, I considered a DALLE-3 image i depicting concept class j, and obtained its embedding $e_{j,i}$. I then computed the cosine similarity between p_j and $e_{j,i}$, and between p_j and all the exemplars from other concept classes $k \neq j$. Finally, I computed the difference between these cosine similarities. Formally, for every $k \in \{1, \dots, 207\} \setminus \{j\}$ and every $l \in \{1, 2, 3\}$:

$$d_{k,l} = cos(p_j, e_{j,i}) - cos_{k \neq j,l}(p_j, e_{k,l})$$

The resulting difference $d_{k,l}$ quantifies the differential similarity of the appropriate and inappropriate exemplars, $e_{j,i}$ and $e_{k,l}$, to the prototype p_j. Then, the same procedure was repeated after replacing the real prototype p_j with the predicted prototype $\widehat{p_j}$, namely the embedding of the word describing concept class j mapped from the language space to the vision space:

$$\widehat{d_{k,l}} = cos\left(\widehat{p_j}, e_{j,i}\right) - cos_{k \neq j,l}\left(\widehat{p_j}, e_{k,l}\right)$$

Finally, I computed the Pearson correlation between $\widehat{d_{k,l}}$ and $d_{k,l}$.

b) *Within-class:* I repeated the above procedure but comparing the similarity of an exemplar to its prototype to the similarity of an exemplar *of the same class* to that prototype, that is:

$$d_l = cos(p_j, e_{j,i}) - cos_{l \neq i}(p_j, e_{j,l})$$

Again, I repeated the operation with a predicted prototype in place of the true prototype:

$$\widehat{d_l} = cos\left(\widehat{p_j}, e_{j,i}\right) - cos_{l \neq i}\left(\widehat{p_j}, e_{j,l}\right)$$

And computed the Pearson correlation between $\widehat{d_l}$ and d_l.

The correlation coefficients of a) and b) were summed, and the configuration of α and λ parameters leading to the highest combined correlation was selected (table 3.1).

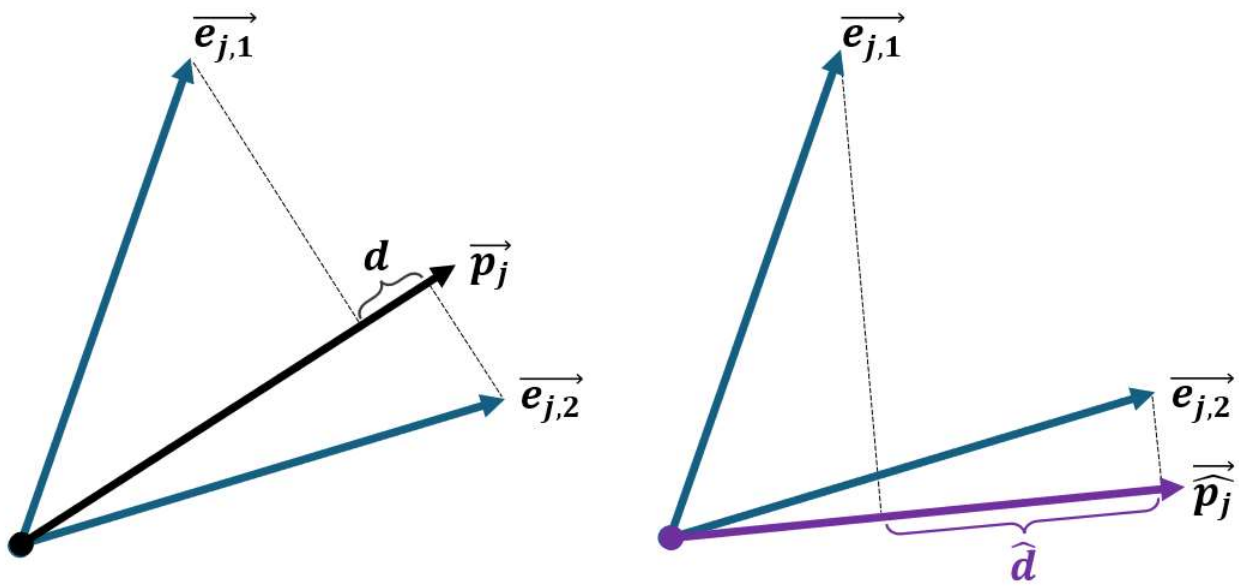

Figure 3.4: (left) The embeddings of exemplar 1 and 2 of concept j, $e_{1,j}$ and $e_{2,j}$, are compared to their concept class prototype p_j via cosine similarity, and d is the difference between their cosine similarities. (right) The same operation is repeated using the predicted prototype $\widehat{p_j}$, resulting in the predicted difference $\hat{d}$. We selected the language-to-vision mapping parameters based on the correlation between $\widehat{p_j}$ and p_j.

Table 3.1: Results of the language-to-vision mapping on the development and test sets for all three CNNs. All within- and between-class correlations were significant at $p < .0001$ (except within-class correlation of VGG-F-1024, significant at $p = .0005$). Significance was assessed by randomly permuting the exemplar label 10000 times and re-computing the Pearson correlations.

CNN; layer	α	λ	**comparison type**	**Development set (Pearson r)**	**Test set (Pearson r)**

VGG-M-1024; fc7	0.001	0.7068	*between-class*	0.47	0.34
			within-class	0.38	0.19
GoogleNet; 5a_3x3_reduce	0.1	0.1448	*between-class*	0.60	0.40
			within-class	0.61	0.37
DarkNet-53; conv50	0.001	0.1501	*between-class*	0.52	0.42
			within-class	0.65	0.50

At testing, the mapping model thus selected was used to map CAOSS embeddings from the language space to the vision space, and the corresponding between- and within-class correlations were computed and used to select the CNN. The best-performing CNN was DarkNet-53, with between- and within-class correlations of $r = 0.42$ and $r = 0.50$, respectively. Note that the compounds used for testing, while attested in English, were absent in the training set of the CNNs models employed, Moreover, their language embeddings were derived from the CAOSS model rather than word2vec, and their images were generated by DALLE-3. Thus, results provided an empirical basis for the claim that a language-to-vision mapping preserves the relationship between prototypes and exemplars *of novel combinations*, which can be used to recover the class and typicality of the exemplars.

3.2.4 Experiment

Participants

A Power analysis was conducted to estimate the sample size required to observe the effects of interest. To do so, I relied on typicality ratings obtained for 1500 existing concepts in Günther et al., (2022). For all 1487 concept categories for which ImageNet-21k were available, predicted vision prototypes were generated using the same mapping function, language and vision spaces employed in the main study (note that, because concepts were mostly made of single words, word2vec embeddings were used rather than CAOSS embeddings). Focusing on the most and least typical images for each category, the effect size was estimated with a generalized linear mixed model identical to analyses 1 and 2 (section 3.3.2 and 3.3.3). The simr package (Green & Macleod, 2016) was used to estimate the required sample size, opting for a sample size of N = 160 which, using simr powerSim function, resulted in a statistical Power of 96.7% for the within-interpretation condition and of 90.2% for the between-interpretation condition. Finally, because Günther et al.' (2022) rating

task was untimed, the effect size for selection times could not be estimated with this method. However, with a sample size of 160, 40 trials per participant and 2 conditions, the resulting 6400 datapoints are above the general recommendations (N = 3200) of Brysbaert, (2018) for linear mixed models analyses of reaction time data.

The Prolific platform was used to recruit adult native English speakers located in the US with normal or corrected-to-normal vision. Experiment duration was estimated to be 20 minutes, and participants were reimbursed $3,5 for their participation ($10,5 per hour). 165 participants were initially recruited, anticipating a small dropout rate.

Materials

Novel combinations: 80 novel compounds were created manually, respecting the following criteria:

i. No two novel compounds have identical head constituents;
ii. Head and modifier constituents are concrete words (concreteness rating of at least 4, following Brysbaert et al., (2014) norms, in line with Hebart et al., 2019);
iii. Head and modifier constituents have a raw frequency of at least 50 in Baroni et al., (2014) corpus, to ensure stability of their word2vec embeddings;
iv. Novel compounds appear at most once in the Baroni et al., (2014) corpus (combined frequency of closed, hyphenated and open forms), ensuring that it is extremely unlikely that any participant might have encountered them.

Moreover, compounds were created such that both destructive and non-destructive interpretations could be represented visually, regardless of the interpretation's plausibility, while striving to avoid compounds that might be interpreted metaphorically. I also tried to span a continuum of "interpretation bias" such that, on one end of the spectrum, a compound would have a bias towards the destructive interpretation (e.g., "leopard blanket") and a bias towards the non-destructive interpretation on the other end of the spectrum (e.g., "walnut bread"), with combinations equally amenable to both interpretations in the middle (e.g., "snow duck"). These considerations served the purpose of obtaining stimuli that could be interpreted in different ways and with various degrees of certainty (that is, to avoid the case in which the same interpretation is preferred for all stimuli).

Synthetic images: The objective was to obtain 5 images for each interpretation of each novel compound. 15 prompts for each interpretation of each compound were manually written, and their synthetic images were generated with DALLE-3 (style = "natural", image size = 1024×1024). After

image generation, 5 images were selected following the same criteria outlined for validation concepts. Image generation was repeated until at least 5 images were deemed appropriate. Thus, in this first step, a set of 80 novel compounds was created, each compound defined by a destructive (D) and a non-destructive (ND) interpretation, and each interpretation depicted by 5 images. Then, the mapping model was used to generate predicted vision prototypes of each novel compound. For D and ND separately, the closest (c) and farthest (f) image from the prototype were selected. Finally, a synthetic image for the head concept was selected randomly from the validation set. Thus, each novel compound was depicted by one image for each of the following types (figure 3.5):

i. Destructive interpretation, close (Dc);
ii. Destructive interpretation, far (Df);
iii. Non-destructive interpretation, close (NDc);
iv. Non-destructive interpretation, far (NDf);
v. Head concept (H);

Figure 3.5: Examples of the synthetic images used in the experiment.

Human judgements were collected via a forced choice task whereby participants selected one of two images presented on the screen (figure 3.7). Thus, the following sets of image pairs were created to address three hypotheses (see Introduction):

1) *Between interpretations condition:* Given a novel compound, do participants select an image depicting its destructive or non-destructive interpretation? To cover this case, image pairs depicting different interpretations were considered (Figure 3.6; blue arrows)[8].
2) *Within interpretation condition:* Given two images depicting the same interpretation of a novel compound, which image is preferred (i.e., what is the most prototypical depiction of the concept)? Pairs where images depicted the same interpretation were considered (Figure 3.6; orange arrows).
3) *Selection speed:* Can we predict the speed at which an image depicting a novel compound is chosen? Pairs where one image depicted a novel compound (in either interpretations) and the other depicted the head concept (Figure 3.6; gray arrows) were considered.

[8] This case admits 4 possible comparisons, namely Dc-NDc, Dc-NDf, Df-NDc, and Df-NDf. Because comparisons are based on distances in semantic space, including pairs that share an exemplar would lead to a violation of statistical independence; thus, we only considered 2 pairs: Dc-NDc and Df-NDf. Note that this issue does not apply to the selection speed condition (Figure 3.6; gray arrows): because the Dc, Df, NDc, and NDf exemplars are compared to the same representation (the compound's head, H), including a random intercept for compound identity effectively accounts for such clustering.

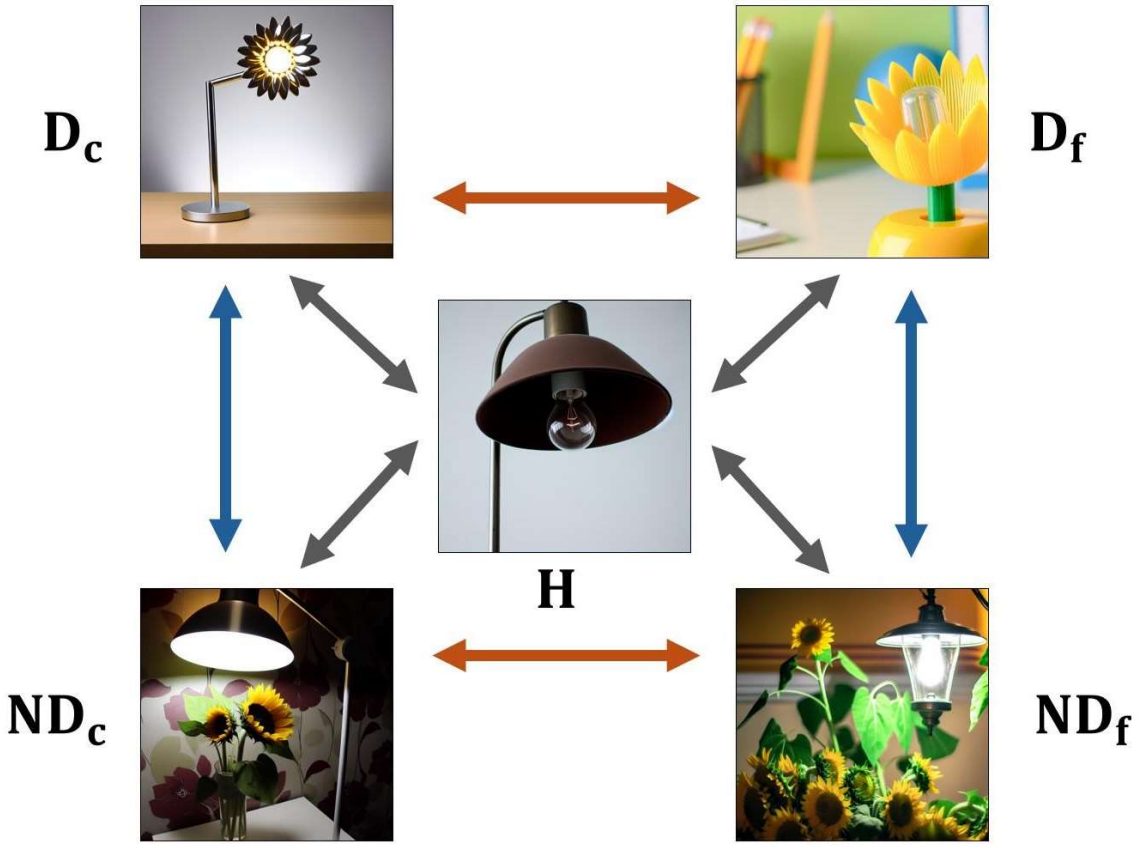

Figure 3.6: Examples of image pairs for forced choice task trials involving the "sunflower lamp" compound. Blue arrows correspond to between-interpretations comparisons (Dc-NDc, Df-NDf), orange arrows correspond to within-interpretations comparisons (Dc-Df, NDc-NDf), and gray arrows correspond to baseline comparisons with the head concept (Dc-H, Df-H, NDc-H, and NDf-H), which were used to test effects on selection speed.

Therefore, each of the 80 novel compounds was characterized by 8 image pairs (the arrows in Figure 3.6), for a total of 640 image pairs. Because the goal was to predict the visual appearance of *novel* (and hence *unseen*) combinations, each participant could not be exposed to the same compound twice. Therefore, stimuli were divided into 8 lists of 80 image pairs, where each image pair depicted a novel compound, and no novel compound was presented twice in the same list. Across lists, a compound was never described by the same image pair twice. Thus, each list comprised 20 trials belonging to the between-interpretation condition, 20 trials belonging to the within-interpretation condition, and 40 trials belonging to the selection speed condition. In addition, 20 trials of (non-repeating) familiar compounds were included, each defined by two synthetic images, one depicting the compound, the other the head constituent. Because the ground truth of familiar compounds' visual appearance is known, these stimuli were used to exclude participants with low accuracy. Familiar compound trials were the same in all lists. Compounds were assigned to the 8 lists randomly, and the position of the predicted image was randomized, ensuring that half appeared on the left side and half on the right

side. Finally, to control for image location, a copy of each list was created whereby the position of all images was flipped (left-right), resulting in 16 lists. During data collection, participants were randomly assigned to one of the 16 lists, and the trial order of stimuli presentation was randomized.

Procedure

After giving their consent, participants were provided with the instructions. Participants were told that they would be presented with words referring to familiar concepts (for example, "ice cream") or to novel concepts (for example, "fire eel"), and were instructed to imagine how the concept should look like, focusing on the image that came to their mind. They were told that they would then see two images, and that they would have to select the one that most closely resembles their intuition by pressing, as fast as possible, the "A" or "L" keys, indicating the selection of the left or right image, respectively. Participants were explicitly told that, in some cases, both images would depict the same concept and that they should nonetheless select the one that best matched their intuition.

Each experimental trial comprised the following steps (Figure 3.7). First, a compound word was displayed in uppercase letters and written in spaced format in the center of the screen. The word remained on the screen for 1 sec, followed by a blank screen (3 sec) and a fixation cross (0.5 sec). Then, two images were displayed next to one another. The images remained displayed until either the "A" or "L" keys were pressed. Then, participants saw a blank screen (1 sec) and a fixation cross (0.5 sec) before beginning the next trial. Participants performed 10 practice trials to familiarize with the task. During the image presentation phases of the practice trials, information on which response keys to use remained on the screen. The practice comprised 5 attested compounds and 5 novel compounds, none of which belonged to the item set presented in the actual experiment. At the second erroneous response for an attested compound, the experiment ended. Otherwise, participants progressed to the experiment, comprising 100 trials. At the end of the 100 forced-choice trials, participants completed the Vividness of Visual Imagery Questionnaire (VVIQ; Marks, 1973), after which the study ended. The VVIQ is a 16 item questionnaire used to identify people with "aphantasia", i.e., reduced or absent visual imagery (Zeman et al., 2015, 2020). Because participants falling in this category might rely on different cognitive processes to carry out the main task (Logie, 2018), VVIQ scores were used to identify and exclude them from statistical analyses.

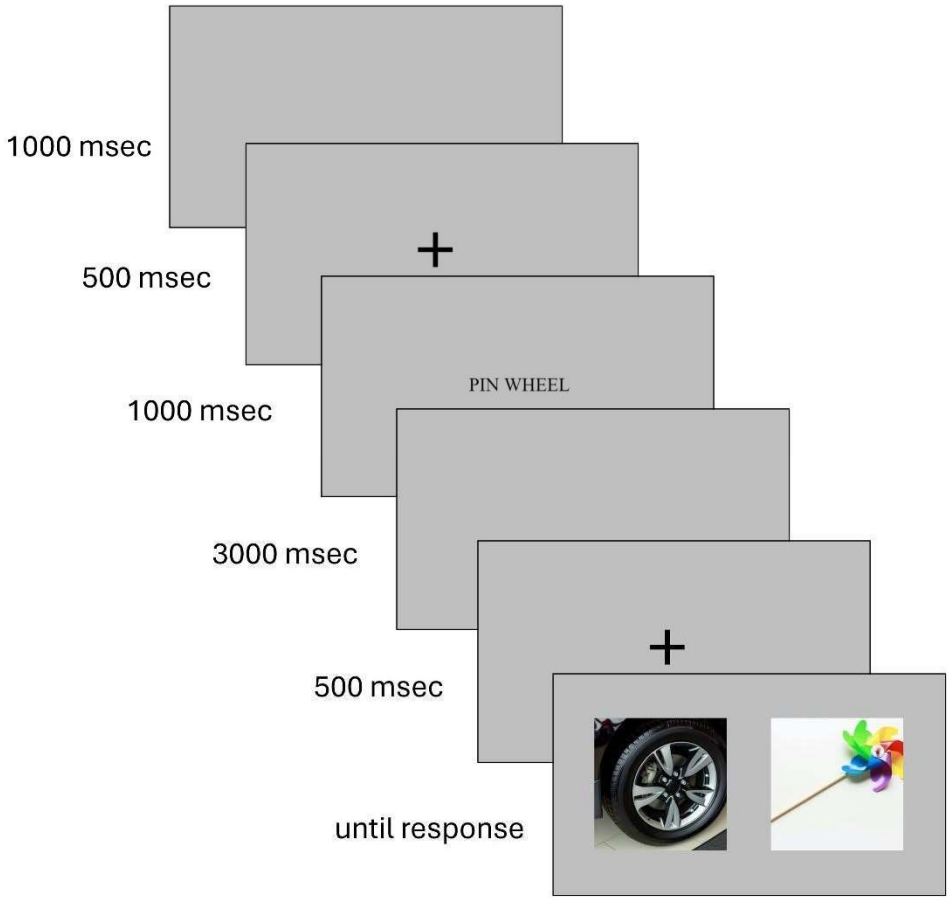

Figure 3.7: The structure of an experiment trial.

3.3 Results

3.3.1 Descriptive results

Data was collected from 165 participants. 5 participants were excluded due to low accuracy (i.e., more than 6 errors when judging the 20 familiar compounds). While I expected 2-3% of the sample to meet the threshold for aphantasia (VVIQ < 24), the lowest VVIQ score observed was 26 (Appendix A.1.2). Then, separately for D and ND, I inspected the percentage of compound images that were chosen over the head image. If, for a given compound, both D and ND images were selected less times than the head image, one would take it as an indication of a poorly designed stimulus, and drop it from any subsequent analyses. While none of those cases were observed, some images depicting the destructive interpretation of a compound were indeed selected less times than the head image (namely “walnut bread”, “bark roach”, “palm bird”, and “ash hourglass”; Figure 3.8). Note that preferring the image of the head concept does not necessarily imply that the image or the interpretation of the compound in question is flawed. Since participants were asked to anticipate how

the compound would *look like*, visual features might have played a larger role in participants' judgment than conceptual ones. Taking the most extreme case in the sample, "walnut bread", it is possible that, when forming the mental image of "a type of bread with walnut pieces over it", the imagined bread might resemble more the head concept (a normal bread) than the destructive interpretation (a type of bread shaped like a walnut).

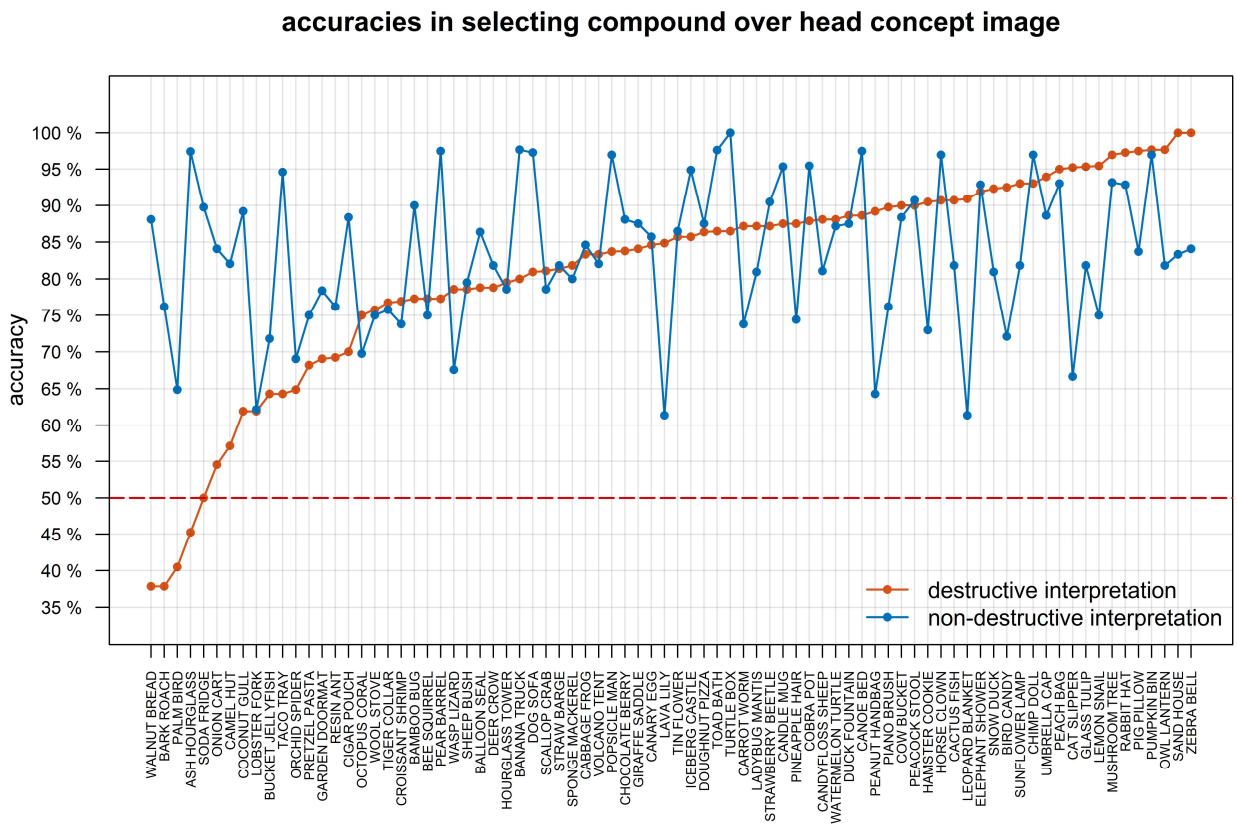

Figure 3.8: Proportion of times the image of the destructive or non-destructive interpretation of a compound was selected over the image of the head concept.

Notably, the actual preferences for either interpretation were remarkably close to my objective: as figure 3.9 shows, participants' preferences span a continuum, from a strong preference for D, to a strong preference for ND, passing through an ambiguous midpoint. A slight overall preference for destructive interpretations was also observed: D were selected more often than ND for 47 compounds, while the reverse was true in 31 cases, and the two were selected an equal amount of times in the remaining 2 cases.

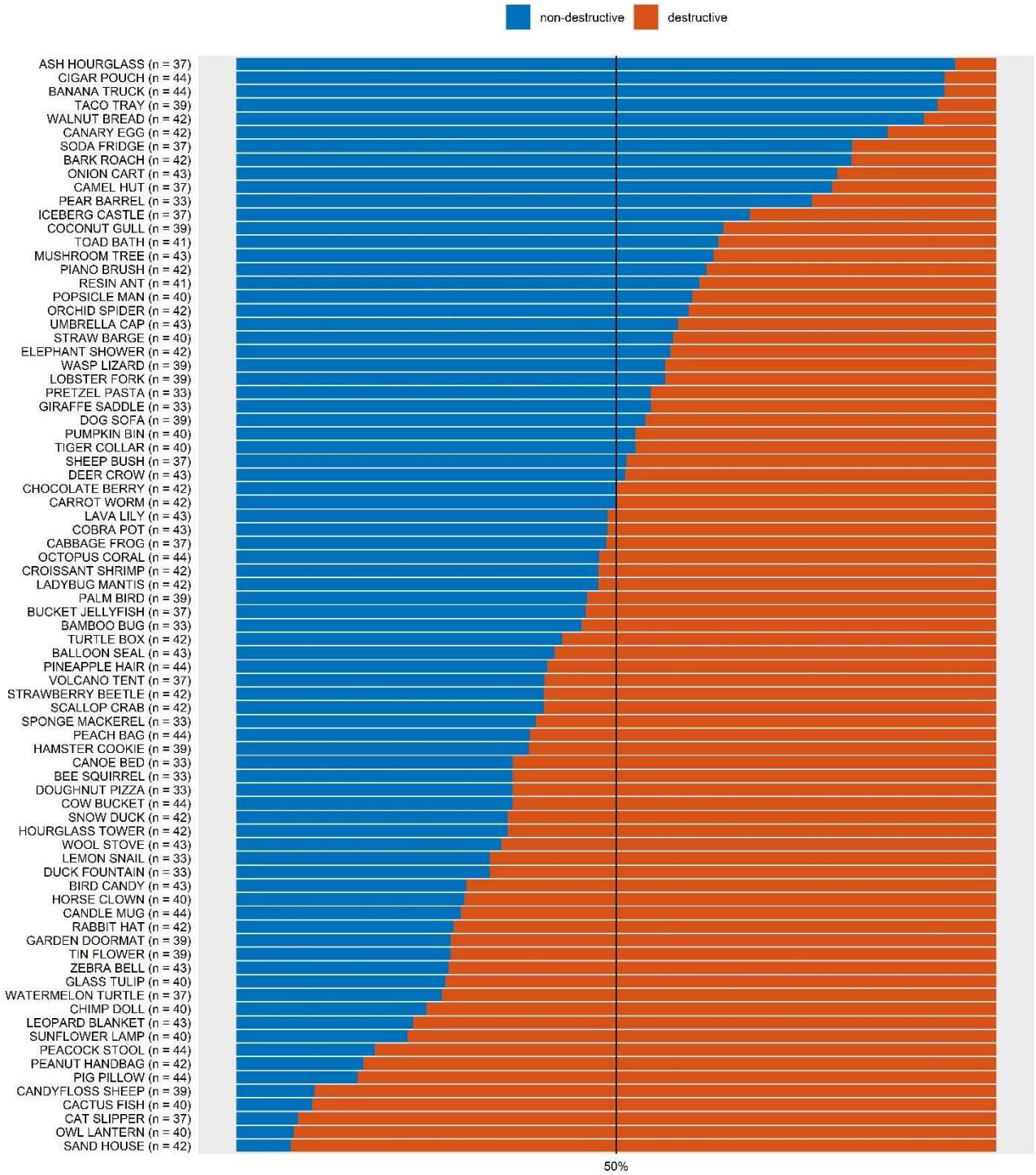

Figure 3.9: Proportion of times the destructive interpretation was preferred over the non-destructive one, computed from all trials where the forced choice selection was between Dc and NDc, or Df and NDf (the number of trials considered is displayed in parentheses).

3.3.2 Analysis 1: Predicting interpretation type

Exclusion criteria: Trials lasting less than 150 msec were excluded, resulting in 3197 datapoints.

Statistical model: To test hypothesis 1, I defined a generalized linear mixed model with a Binomial link function. The dependent variable was a dichotomous variable coding for whether the model-predicted exemplar was selected (1) or not (0). The continuous predictor was the difference between the cosine similarity of the closest exemplar and the prototype, and the farthest exemplar and the prototype. This metric, which I call *differential similarity*, should capture the increased adequacy/plausibility of the closest exemplar over the farthest. Random intercepts for compounds and participants were included, resulting in the following statistical model:

$$choice_i \sim \left(cos\,(e_{close_i}, \widehat{p}_i) - cos\,(e_{far_i}, \widehat{p}_i)\right) + (1|compound) + (1|participant)$$

Where $\widehat{p}_i$ is the embedding of the predicted vision prototype of compound i, e_{close_i} the embedding of the closer exemplar image, and e_{far_i} the embedding of the farther exemplar image, while $choice_i$ codes for whether the closer image was chosen ($choice_i = 1$) or not ($choice_i = 0$). Here, the e_{close_i} and e_{far_i} exemplar embeddings belong to different interpretations. The DHARMa package (Hartig, 2020) was used to ensure that model assumptions were met. Finally, model criticism was performed by excluding datapoints with residuals[9] of magnitude greater than |2.5| and re-fitting the model on the resulting dataset. A simple Binomial test was also conducted to assess whether the number of accurate responses exceeds that of errors, regardless of differential similarity.

Results: The probability of an accurate selection (52.5%) was significant according to a Binomial test ($p = .005$). The random intercept for participants' identity was removed, as it led to singular fit (Variance $< 10^{-10}$). In the resulting logistic mixed model, differential similarity was significant ($\beta = 5.09, p = .015$), indicating that accuracy was modulated by differential similarity: the closer an image was to the language-predicted representation, compared to the alternative, the more often its was selected (Figure 3.10). Results were robust to model criticism ($\beta = 5.15, p = .015$).

[9] Residuals were estimated via simulation with the DHARMa package, using the simulateResiduals functions and 1000 simulations.

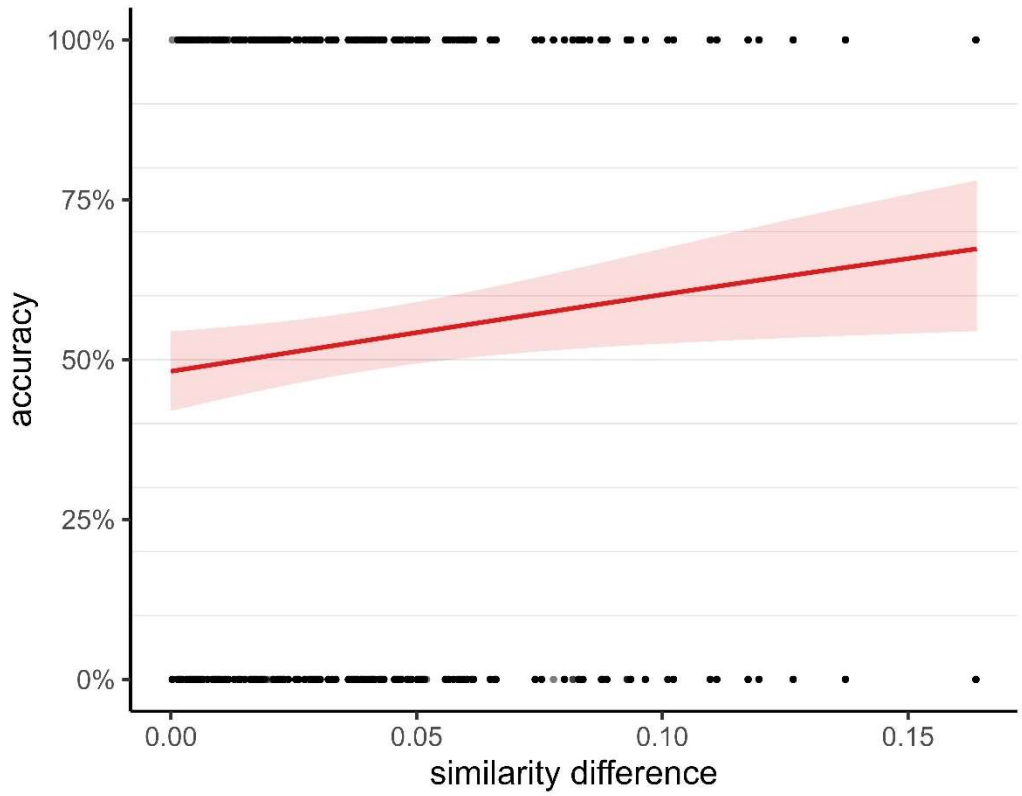

Figure 3.10: Fixed effect of differential similarity on selection accuracy (i.e., whether the predicted image was selected or not).

3.3.3 Analysis 2: Predicting typicality

Exclusion criteria: Trials lasting less than 150 msec were excluded, resulting in 3200 datapoints.

Statistical model: The statistical model and procedure are identical to analysis 1. Here, however, the e_{close_i} and e_{far_i} exemplar embeddings refer to the same interpretation, either D or ND. Because there were no theoretical reasons to assume different effects of the predictor on D and ND, data for both D and ND trials were aggregated.

Results: According to a Binomial test, the probability of accurate selection was not significant (48.3%, $p = .063$), as was the effect of differential similarity estimated by the logistic mixed model ($\beta = 2.40, p = .109$).

3.3.4 Analysis 3: Predicting selection times

Exclusion criteria: Trials lasting less than 150 msec and more than 4000 msec were excluded (see Appendix A.1.4 for a robustness analysis varying the reaction times thresholds), resulting in 6042 datapoints.

Statistical model: I tested hypothesis 3 with a linear mixed model. Log-transformed reaction times were included as the dependent variable, while the main continuous predictor was the difference between the cosine similarity of the exemplar (Dc, Df, NDc, or NDf) and the prototype, and the head concept and the prototype. Moreover, another predictor coded for the compound interpretation type (ND = 1, D = 0). For this analysis, I considered only those trials where the head image was *not* selected. Again, participants and compounds were included as random intercepts:

$$log\ (RT)_i \sim (cos(e_i, \widehat{p_i}) - cos(head_i, \hat{p})) + interpretation + (1|compound) + (1|participant)$$

Finally, model criticism was performed by excluding residuals greater than |2.5| and re-fitting the model on the truncated dataset.

Results: Differential semantic similarity significantly predicted log-transformed reaction times ($\beta = -0.24, p = .011$), a result that remained significant after model criticism ($\beta = -0.22, p = .011$). As hypothesized, the more similar a compound word image was to the predicted prototype compared to the head image, the faster its selection (Figure 3.11).

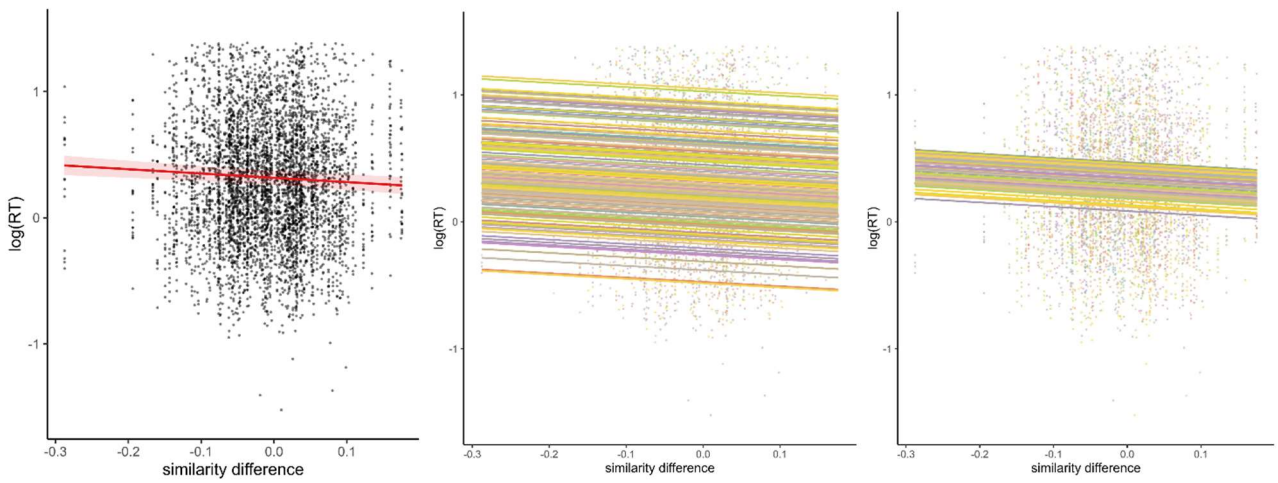

Figure 3.11: Analysis 3 model estimates. (left) Fixed effect of differential similarity on log-transformed reading times. (middle) Random intercepts for participant identity (right) Random intercepts for word identity.

3.3.5 Exploratory analyses

Images with more similar representations are harder to discriminate. Since differential similarity (the difference between two image-prototype similarities) and image similarity (the similarity between two images) are correlated ($r = 0.18, p < .0001$), it's possible that the results observed on selection speed could be explained or dampened by mere image similarity. Thus, the cosine similarity between the two images was included as a predictor in the statistical model. This addition improved the model's Akaike Information Criterion ($\Delta AIC = 3.2$) but, while differential similarity remained significant ($\beta = -0.20, p = .037$), image similarity was not ($\beta = -0.11, p = .136$), indicating that the latter did not play a role in predicting response times.

While the CAOSS model was used to generate combinatorial representations, it is possible that image selection times could be predicted just as well, or better, with simpler, non-combinatorial models. Thus, three control models were defined where the predicted prototypes were generated from the embeddings of i) the modifier constituent, ii) the head constituent, or iii) the unweighted sum of modifier and head constituents (i.e., additive model; Mitchell & Lapata, 2010). Each control model was tested in a separate linear mixed model identical to that of Analyses 1 and 3, and models were compared based on their AIC. Results show that, for both analysis 1 and analysis 3, language-to-vision mapping based on CAOSS embeddings did not lead to the best model fit; instead, the head concept provided the best fit, followed by CAOSS, by the additive model and, finally, by the modifier constituent (Table 3.2; see Appendix A.1.5 for additional details about model comparison). Because the head constituent proved to be the best model of participants' behavior in analyses 1 and 3, the same model comparison was also carried out for analysis 2. Again, the model including predictions from the head constituent performed best, followed by the additive model, by the CAOSS model, and by the modifier constituent (Table 3.2). Notably, differential similarity based on the head constituent was highly significant in predicting image typicality ($\beta = 3.50, p = .0002$)[10].

Table 3.2: Results of language model comparison. ΔAIC is the difference in the Akaike information criterion between a given model and the best-fitting model (i.e., the model with the lowest AIC).

model	ΔAIC		
	analysis 1 (between-interpretations)	**analysis 2** (within-interpretation)	**analysis 3** (speed of selection)

[10] Differential similarity based on the head constituent was also highly significant in analysis 1 ($\beta = 3.56, p = .0006$) and in analysis 3 ($\beta = -0.34, p = .0001$).

CAOSS $M\overrightarrow{w_1} + H\overrightarrow{w_2}$	5.9	11.3	8.1
Modifier constituent $\overrightarrow{w_1}$	11.5	13.0	14.8
Head constituent $\overrightarrow{w_2}$	0	0	0
Additive model $\overrightarrow{w_1} + \overrightarrow{w_2}$	8.0	10.6	11.5

Analysis 3 focused on all the trials where an image depicting a compound was chosen over the image of its head constituent. Then, I tested whether differential similarity could also predict the response times in trials from the within- and between-interpretation conditions. I expected two properties to affect selection times, namely the absolute similarity of the chosen image to the predicted prototype, and the competition exerted by the other. Because, in Analysis 3, all images of a compound were compared to the same head image (figure 3.6; gray arrows), random intercepts for compounds could account for the former. The same does not hold for within- and between-interpretation conditions, where images belong to different pairs (figure 3.6; orange and blue arrows). Thus, linear mixed models included predictors coding for the similarity between the chosen image and the predicted prototype ($cos\ (e_{i_1}, \widehat{p_i})$), the similarity between the non-chosen image and the predicted prototype ($cos\ (e_{i_0}, \widehat{p_i})$) and, in the case of the within-interpretation condition, a predictor coding for interpretation type (D vs. ND)[11]. Again, random intercepts for compounds and participants were included:

$$log\ (RT)_i \sim cos\left(e_{i_1}, \widehat{p_i}\right) + cos\left(e_{i_0}, \widehat{p_i}\right) + interpretation + (1|compound) + (1|participant)$$

Following analysis 3, trials faster than 150 msec and slower than 4000 msec were excluded.

[11] For comparability, the same analytic setting was also applied to Analysis 3 (i.e., compound-head pairs where the compound image was selected). The similarity of the chosen image (i.e., the compound image) and the predicted prototype was highly significant ($\beta = -0.26, p = .008$), but not the similarity of the non-chosen image (i.e., the head image) and the predicted prototype ($\beta = 0.10, p = .609$). Again, the head concept proved to be the best-fitting model ($\Delta AIC_{CAOSS-hea} = 12.8$, $\Delta AIC_{modifier-head} = 14.2$, $\Delta AIC_{additive-head} = 17.1$).

(i) *Between-interpretations:* Neither the similarity of the chosen and not chosen image to the prototype were significant predictors of reaction times ($\beta = -0.16, p = .28$; $\beta = -0.04, p = .77$, respectively). Results remained non-significant when the predicted prototype was generated from the head constituent in place of the CAOSS model ($\beta = -0.04, p = .72$; $\beta = -0.09, p = .38$).

(ii) *Within-interpretations:* Again, neither the similarity of the chosen and the non-chosen image to the predicted prototype were significant ($\beta = -0.02, p = .82$; $\beta = 0.15, p = .18$, respectively). Results were also non-significant when the head constituent was used as the language model, in place of CAOSS ($\beta = -0.12, p = .21$; $\beta = 0.18, p = .06$).

3.4 Discussion

Knowing that constituents can be linearly combined in language space by the CAOSS model (section 1.2.2; Chapter 2), and knowing that language and vision spaces can be related via a simple linear mapping, Chapter 3 tested the possibility that the visual representation of novel combined concepts can be obtained by the sequential application of language composition and language-to-vision mapping. Specifically, Chapter 3 tested the hypotheses that language-to-vision mapping can predict (1) which interpretation of a novel compound is preferred, (2) which image depicting the same interpretation is preferred, (3) the time taken to select the preferred image.

Predicting interpretation type: Hypothesis 1 was confirmed: differential similarity predicts whether participants preferred a destructive or a non-destructive interpretation of a novel compound, consistent with the notion that language distributional information can be leveraged to predict extralinguistic information, and extending it to the case of novel concepts. More specifically, the more an image was similar to a language prediction compared to the alternative, the more frequently it was selected. However, contrary to my expectations, simply relying on the head constituent provides a sizeable advantage in model fit over the CAOSS model, suggesting that participants' choices of D vs. ND interpretations are best explained by a simpler model that leverages the distributional information of the head (the semantic core of the compound), rather than the full combinatorial meaning of the compound.

Predicting image typicality: When based on predictions from the CAOSS model, differential similarity does not predict the image typicality of novel compound words, contrary to hypothesis 2. However, when predicted prototypes were generated based on the embeddings of head constituents, differential similarity was significant: the closer an image was to the prototype predicted by the head

constituent, the more likely it was selected over the alternative. Here, images provided different depictions *of the same novel concept,* thus requiring the mapping model to tap into more genuinely unimodal visual information, posing an even harder challenge to a language-to-vision mapping system. Going back to the validation of synthetic images, the results of the Power analysis (section 3.2.4) demonstrate that language-to-vision mapping could indeed predict typicality judgments of familiar and simple concepts from Günther et al. (2022), while in the validation of the mapping model (section 3.2.3), the correlation of 0.50 between true and predicted differential similarities (table 3.1; figure 3.4) further suggests that the typicality of familiar complex concepts can also be recovered via language predictions. In this context, results from analysis 2 extend these findings to the case of complex novel concepts.

Predicting selection speed: Following the preregistered procedure, this analysis was initially focused on the data subset of compound-head image pairs. Because participants were tasked with judging *novel* combinations, it could be assumed that in these trials, even if they were anticipating a different interpretation, participants would at least *not* select the familiar head concept. This made it possible to distinguish between accurate (the compound is selected) and inaccurate (the head concept is selected) trials, and the analyses were conducted on the former subset. Hypothesis 3 was confirmed: the greater the advantage of the selected image – quantified as differential similarity to the predicted prototype – the faster its selection. Moreover, even in this condition, the head concept led to the best model of participants' behavior, outperforming CAOSS. Finally, when the statistical model included both the similarity of the chosen and non-chosen (head) image to the predicted prototype, only the former was significant, indicating that language predictions appeared not to influence rejection times of the incorrect (head) image. However, these findings did not generalize to subsequent analyses targeting the other trial types (i.e., between- and within-interpretation conditions). The ECCo model does not propose a strict sequential engagement of the simulation and language systems; rather, "the two systems continually feed into one another (i.e., words help to activate simulations, and simulations help to activate words)." (Lynott & Connel, 2010, p. 4). While it is impossible to address how these two systems might have evolved in time, results at least appear inconsistent with a situation where distributional information guides or "pre-activates" visual representations. If that were the case, one would expect to observe not only a facilitatory effect of prototype-image similarity for the chosen image (which is indeed present) but also an inhibitory effect of prototype-image similarity for the other, non-chosen image. The lack of results in the between- and within-interpretation subsets also speaks against the idea of preactivation.

Why did the effect of language prediction on selection times emerge only in the (accurate) compound-head trials? One possibility is that the influence of language prediction may depend on the degree of interpretation specificity required to carry out the forced-choice task. Indeed, in principle, participants could have responded to head-compound pairs by simply selecting the image *other than the head*. Thus, the reduced demands of compound-head pairs might have made it possible for effects of language to emerge, consistent with the ECCo model (Lynott & Connell, 2010; Connell & Lynott, 2011), which posits that language distributional information typically provides a "shallower", coarse-grained approximation of the combined concept, which would usually require sensorimotor simulation to be generated fully. Indeed, because the head constituent provided the best model fit, similarity to the predicted prototype can be interpreted as the degree to which the head concept is recognizable in the image of the compound, without calling for more complex processes of conceptual combination. For example, when responding to "cactus fish", participants might have ruled out the image of a fish (the head), while the time taken to select the other image would have been proportional to the easiness of ascertaining that it was, indeed, depicting "a type of fish". Here, while analyses 1 and 2 prove that differential similarity *did* predict image selection in these conditions, the fact that it did not predict selection times may indicate that additional processing was carried out on top of language prediction, dampening its effects. Consistent with this possibility, reaction times in compound-head trials were significantly lower than those of between- and within-interpretation trials (see Appendix A.1.6). In other words, participants choices were consistent with language predictions, but the latter's effect on decision times might have been observable only when other, possibly deeper processes of conceptual combination were not engaged, i.e., in the "easier" head-compound trials.

Model comparison: Interestingly, across analyses, the rank of language models remained almost invariant, with the head concept leading to the best-performing statistical models, followed by CAOSS and the additive model, and finally by the modifier constituent (Table 3.2). This result runs counter to theories like CARIN and RICE (Gagné & Shoben, 1997; Spalding, Gagné, Mullaly & Ji, 2010), positing a primary role of the modifier in determining the interpretation of a compound, as well as empirical results from the compositional distributional semantics tradition, reporting effects driven by the properties of the modifier (Günther et al., 2020; Günther & Marelli, 2018, 2020; Marelli et al., 2017). Indeed, according to the ECCo model (Lynott & Connell, 2010) and consistent with the present study's stimuli, it is usually the modifier that is either reduced or left intact, while the head typically remains unaltered. This fact might explain why models based on the modifier constituent performed so poorly: the meaning of a modifier as a standalone word is substantially altered when used in destructive combinations, and is thus less reliable. The problem persists in the additive model, which uses the same representation of the standalone modifier, while CAOSS might have not been

able to provide an adequate representation of the (transformed) modifier, which, when added to the head representation, might have led to noisier estimates inferior to those of the standalone head concept.

Overall, results are consistent with the reported accuracy of (linear) language-to-vision mapping (Günther, Petilli, Vergallito, et al., 2020; Lazaridou et al., 2014; Li et al., 2023; Merullo et al., 2022; Roads & Love, 2020; Sorscher et al., 2022), and extend these findings from the case of simple words to that of novel complex words. In this sense, while it has been argued that language distributional information provides only a coarse-grained pointer to sensorimotor experience (Lowerse, 2018), insufficiently precise to characterize complex combinations (Lynott & Connell, 2010; Connell & Lynott, 2011), this Chapter provides evidence of the opposite: differential similarity was positively related to the likelihood of selecting a given interpretation of a novel compound and a specific image that best depicted such interpretation. However, the consistently superior performance of the head suggests that results are best accounted for by a mapping system that does not rely on constituent combination but is based, more simply, on the distributional meaning of the head concept, the semantic and syntactic core of the compound. This fact, coupled with the small effects on accuracy and the lack of effects of differential similarity on reaction times in the between- and within-interpretation conditions, suggests that, in line with the ECCo model, distributional information was only recruited to approximate compound meaning, which likely involved additional combinatorial processes to be generated (Lynott & Connell, 2010). In this sense, developing a compositional model that directly operates over vision embeddings could constitute a promising development of the present Chapter. Finally, these findings complement those of Günther et al., (2020), who showed that visual combination affects processing times in purely linguistic tasks (naming, lexical decision, timed sensibility judgments). Together, these two studies suggest that both language-based and vision-based combinations are enacted automatically, irrespective of (modality-specific) task demands. Finally, it is worth noting that the time between the presentation of written compounds and their images was relatively long (i.e., 3500 msec), ensuring that participants could form detailed mental images, but possibly dampening fast language predictions (Lynott & Connell, 2010). Thus, future studies might reduce the stimulus onset asynchrony, possibly revealing additional signatures of language combination on chronometric behavior.

Chapter 4

The neural bases of semantic composition: A model-based fMRI re-analysis

4.1 Introduction

Neuroimaging studies have identified a set of candidate brain regions recruited for simple forms of semantic combination in noun-noun and adjective-noun phrases, most of which lie within the frontal, temporal and parietal regions comprising the general semantic network (Binder et al., 2009; Frankland & Greene, 2020; Jackson, 2021). Among these, the left inferior frontal gyrus (LIFG), the left and right angular gyrus (LAG, RAG), and the left anterior temporal lobe (LATL) have been consistently observed and related to specific processes (Boylan et al., 2015, 2017), leading to claims about the type of semantic combination they support (Coutanche et al., 2018; see general introduction, section 1.2.2).

While these studies have provided a rich picture of the neural bases of semantic composition, most have looked at changes in single-voxel signal resulting from the manipulation of high-level properties (e.g., plausibility, familiarity, literalness), which, through appropriate contrasts (e.g., complex pseudowords vs. complex words), are related to semantic composition. Thus, this approach focuses on isolating composition-specific brain activity, rather than characterizing the underlying transformations. For example, contrasting interpretable and uninterpretable word pairs can reveal something about whether and where composition is accomplished, but does not shed light on *how* combination is actually carried out. Second, while contrasts can code for different types of composition (see Boylan et al., 2017), such an approach imposes high-level distinctions that might not hold in practice. Indeed, the same stimuli can be processed in alternative ways, both compositional

---

This work was presented at the 2023 edition of the "Workshop on Concepts, Actions, and Objects" (CIMeC, Rovereto, Italy) as the following poster contribution: "Ciapparelli, M., Reverberi, C., & Marelli, M. (2023). Semantic composition in the brain: A representational similarity approach".

This work was also presented at the 2024 "Highlights in the Language Sciences" (Radboud University, Nijmegen, Netherlands) as the following poster contribution: "Ciapparelli, M., Reverberi, C., & Marelli, M. (2024). Neural representations of combinatorial semantics: A model-based fMRI reanalysis".

and non-compositional (Kenett & Thompson-Schill, 2020; Libben, 2014), which the brain might carry out in parallel (Baggio, 2021). In this sense, the neuroscientific study of semantic composition could benefit from shifting its focus from activation to representation. Indeed, from a theoretical standpoint, a system is combinatorial if its symbols can be systematically combined to obtain other symbols; if these symbols carry meaning and if these meanings can themselves be systematically combined to obtain other meanings, then the system is said to be compositional, and the ability to do so is semantic composition (Martin & Baggio, 2020). In the domain of cognitive neuroscience, representations are neural activities that, like symbols, can stand-in for stimuli (Webb, 2006). Therefore, from a cognitive neuroscience perspective, the study of semantic composition can be described as the study of the systematic transformation of meaning-carrying representations. In this context, computational models can provide the formalism to characterize both representations and transformations. A class of models naturally apt to accomplish this task is that of distributional models from the "compositional" tradition (cDSMs; Baroni, 2013; Baroni et al., 2014), whereby word embeddings offer rich representations of constituent meaning and simple algebraic functions specify their compositional transformation. Thus, while previous chapters leveraged representations from the CAOSS model to predict behavior, the current chapter considers multiple cDSMs as candidate representational models of semantic composition in the brain. First, different formalizations of semantic composition can, when interpretable, be related to specific hypotheses of how ROI carry out composition. Second, because different transformations can be applied to the same semantic representations, model comparison can be used to tackle the issue of alternative routes to conceptual combination (such as comparing "stored" and "compositional" representations for the same stimuli; general introduction 1.2.5).

The present chapter consists of a model-driven re-analysis of four fMRI studies, all employing English two-word combinations but varying in task requirements (see table 4.1). Different types of semantic and compositional representations are formalized by (c)DSMs. Consistent with the focus on neural representations, representational similarity analysis (RSA; Kriegeskorte et al., 2008) is used to relate brain and model representations (figure 4.2). In this context, the first and primary goal of the study was to test hypotheses concerning the specific type of semantic composition supported by the four core ROIs of LIFG, LAG, RAG, and LATL. To do so, the following cDSMs were considered (figure 4.1):

- **Additive model** (Mitchell & Lapata, 2010):

$$\overrightarrow{add} = \overrightarrow{w_1} + \overrightarrow{w_2}$$

This model can be taken to capture the simple co-activation of constituents' meanings rather than proper composition, and is included as a baseline. Because core ROIs have been largely implicated in semantic processing, I still expected to find representational similarities between the additive model and the core ROIs.

- **Multiplicative model** (Mitchell & Lapata, 2010):

$$\overrightarrow{mult} = \overrightarrow{w_1} \odot \overrightarrow{w_2}$$

Where $\odot$ represents component-wise vector multiplication, $\overrightarrow{mult_i} = \overrightarrow{w_{1_i}} \times \overrightarrow{w_{2_i}}$. This model is symmetric ($\overrightarrow{w_1} \odot \overrightarrow{w_2} = \overrightarrow{w_2} \odot \overrightarrow{w_1}$) and can be taken to compute the feature intersection of the constituent word vectors (Baroni, 2013). Thus, I expected to observe a representational correspondence between this model and LATL, a similarly non-syntactic feature combiner region (Pylkkänen, 2019) representing the intersective conjunction of concepts (Coutanche et al., 2018; Baron & Osherson, 2011).

- **Compounding as Abstract Operation in Semantic Space (CAOSS) model** (Marelli et al., 2017):

$$\overrightarrow{CAOSS} = \overrightarrow{c_1} + \overrightarrow{c_2} = M \times \overrightarrow{w_1} + H \times \overrightarrow{w_2}$$

CAOSS representations are asymmetric and have been related to relational combinations (Günther & Marelli, 2022; Marelli et al., 2017; Chapter 5). Following the claims and evidence that AG supports relational combinations and thematic relations (Boylan et al., 2017; Coutanche et al., 2018; Schwartz et al., 2011), CAOSS was expected to best match the neural space of LAG and RAG.

Specific hypotheses about the type of composition supported by LIFG are lacking. This ROI was instead considered in relation to the claim that its involvement reflects task difficulty rather than true compositional processing (similar claims related to task difficulty apply to AG). In this context, a significant match with multiplicative or CAOSS models would instead suggest that LIFG does represent compositional information. Moreover, psycholinguistics accounts of compounding posit that constituents are automatically combined (Günther & Marelli, 2020; Libben, 2014; general introduction 1.2.5), a hypothesis compatible with neuroimaging evidence (Liuzzi et al., 2021). To test this claim, confirmatory analyses were re-run in a subset of data where the task did not require explicit semantic processing, expecting to nonetheless replicate the patterns observed in the core ROI analysis.

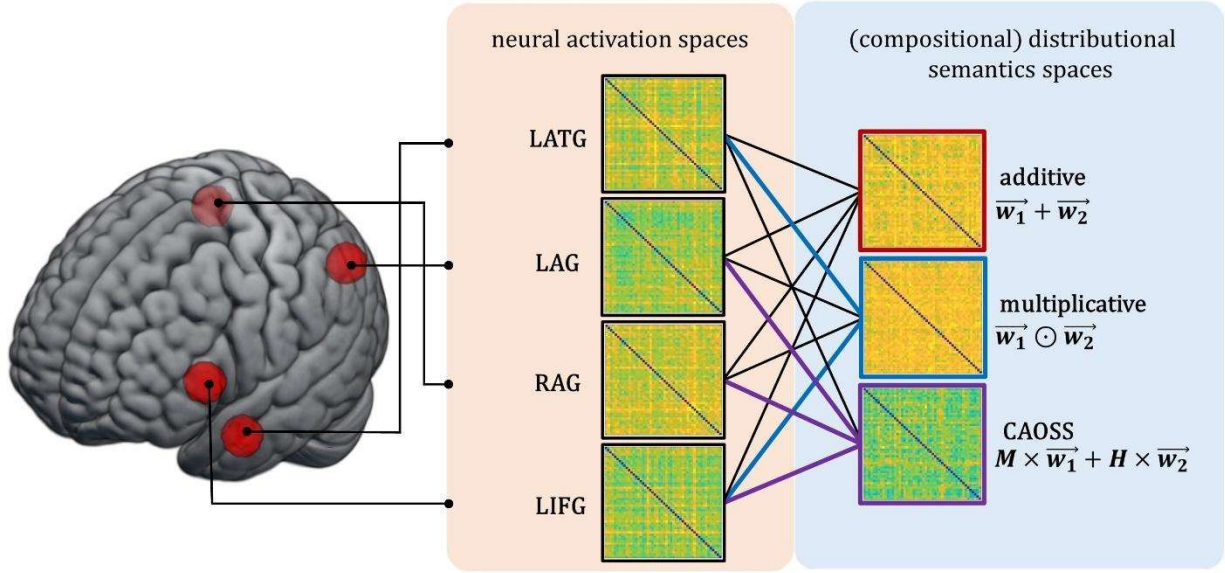

Figure 4.1: cDSM-ROI correspondences tested in confirmatory RSA. The RDMs of the four core ROI (spherical ROIs on the left) are correlated with the RDMs of theoretical models. Colored connections indicate correlations of special interest.

Few neuroimaging studies have used RSA to investigate meaning composition (but see Momenian et al., 2021; Wang et al., 2020), and none with computational models. Thus, exploratory model-driven analyses were also carried out to complement the traditional "activation" perspective, possibly pointing to other regions of interest and, through model comparison, hinting at the specific types of composition they might support. Thus, searchlight RSA were conducted in the broader semantic network. Additionally, two of the four studies re-analyzed include familiar stimuli, which allow comparing non-compositional and compositional representations, dissociating lexicalized meaning retrieval from active semantic composition. Following Libben (2014), both processes should be attempted, resulting in representational similarities from both compositional and non-compositional models (see general introduction 1.2.5).

4.2 Materials and methods

4.2.1 Experimental materials

I searched for fMRI studies that met the following criteria. First, because combinatorial semantics is expressed differently across languages (e.g., compound productivity, headedness; Libben et al., 2020), I decided to focus on a single language, choosing English due to the larger number of fMRI studies available, its compounding productivity (Libben et al., 2020), and the availability of

computational resources and their psycholinguistic evaluation (Baroni, Dinu, et al., 2014; Günther & Marelli, 2020, 2022; Marelli et al., 2017; Mitchell & Lapata, 2010). Therefore, it was required of participants to be native speakers of English, further limiting the search to healthy, adult subjects. Relatedly, I focused on studies that employed English stimuli, either adjective-noun or noun-noun phrases, requiring both constituents to be presented visually and simultaneously. Finally, in order to apply the same data analytic procedures across studies, I considered studies for which raw structural and functional data, either DICOM or un-preprocessed NIfTI files, were available. I contacted the authors of studies meeting the above criteria, and obtained data from four studies: Graves et al., (2010; experiment 1 and 2), Graves et al., (2022), and Wang et al., (2020). Table 4.1 displays the main characteristics of the stimuli, tasks, and participants involved in each study.

Table 4.1: Summary of the stimuli, task, and participant characteristics of the studies included.

study	**stimuli**	**task**	**participants**
Graves et al., (2010; experiment 1)	480 noun-noun phrases, comprising 240 meaningful phrases (e.g., "ski jacket") and their reverse non-meaningful counterparts ("jacket-ski").	1-back task. Participants indicated whether either word of the current phrase matched a word in the same position of the previous phrase (true in 1/6 of trials).	23 participants (13 females); all were healthy adult native English speakers, had normal or corrected-to-normal vision, and were right-handed. Mean age was 24.2 (SD: 3.0).
Graves et al., (2010; experiment 2)	400 noun-noun phrases; 200 meaningful phrases and their reverse non-meaningful counterparts. 200 nonword phrases were used as baseline trials.	Meaningfulness judgements. Participants indicated whether phrases were meaningful, non-meaningful, or made of "nonwords."	22 participants (15 females); all were healthy adult native English speakers, had normal or corrected-to-normal vision, and were right handed. Mean age was 24.7 (SD: 5.4).
Wang et al., (2020)	312 noun-noun phrases. Of these, 96 were place-thing phrases (e.g., "temple-priest"), 96 were category-exemplar phrases (e.g., "dessert-brownie"), 96 whole-part phrases (e.g., "dog-tail") and 24 were unrelated (e.g., "mascara-spoon")	Semantic relation judgements. Participants indicated whether they could think of a relation between the constituents or not. 1/3 of related phrases were primed (i.e., preceded by a phrase with the same relation type)	18 participants (11 females); all were healthy adult native English speakers, and were right handed. Age ranged from 18 to 28 years.
Graves et al., (2022)	192 adjective-noun phrases, varying along the socialness-figurativeness dimensions	Familiarity judgements. Participants indicated whether	Subset of 22 neurotypical participants; all were healthy adult native English speakers,

	(i.e., social figurative, social literal, nonsocial figurative, nonsocial literal)	each phrase they were seeing was familiar or unfamiliar.	had normal or corrected-to-normal vision, and were right handed. Mean age was 21.3.

4.2.2 Imaging methods

Table 4.2 summarizes the characteristics of fMRI data acquisition across the four studies included. The same pre-processing and data analysis pipelines were applied to all studies, accounting for study-specific differences in image acquisition (i.e., TR and acquisition sequence).

fMRI preprocessing

fMRI data preprocessing was carried out with SPM12 software (Wellcome Trust Centre for Neuroimaging, London, UK). Volumes were spatially realigned, unwarped, slice-time corrected and co-registered to structural T1 images. The TsDiffAna toolbox (http://www.fil.ion.ucl.ac.uk/spm/ext/#TSDiffAna) was employed to detect volumes of exceedingly high activation that were potentially due to artifacts. Subsequent analyses were carried out in subject space, with the exception of group-level searchlight RSA, for which volumes were normalized to the standard MNI template.

Univariate analyses

Univariate analyses based on the general linear model (GLM) approach were performed in SPM12. The purpose of these analyses was to obtain a beta map for each phrase stimulus in order to perform subsequent representational similarity analyses (RSAs). In all studies, each phrase was presented once, which required the estimation of beta maps from single trial presentations. The combination of trial-by-trial estimation and rapid event-related designs can lead to unreliable beta estimates due to the collinearity arising from the overlap of BOLD signals from adjacent trials (Mumford et al., 2012). To counter this problem, a staggered estimation approach was implemented. Specifically, beta coefficients were estimated for each trial at intervals of three trials. At the same time, another regressor modeled all remaining trials falling within the trials of interest[12]. Explicitly, three separate GLMs were conducted. The first GLM included predictors for each phrase in trials 1, 4, 7, …, while

[12] This approach follows the rationale of the LS-S method proposed by Mumford et al., (2012), which consists of the estimation of an independent GLM for each trial, where the trial is modeled as the regressor of interest and all the remaining trials are combined into a single regressor. Because of the large number of trials to be modeled, we followed a computationally less demanding strategy.

a single predictor coded for all left-out trials (i.e., trials 2, 3, 5, 6, 8, …). The second and third GLMs included predictors for trials 2, 5, 8, … and 3, 6, 9, …, respectively, and the corresponding left-out trials. In these GLMs, predictors coded for the identity of each phrase or group of phrases against an implicit baseline. Covariates coding for high-level conditions (e.g., whether the word pair was "forward" and meaningful, or "reverse" and non-meaningful; Graves et al., 2010) were not included for two reasons. First, since high-level conditions varied across studies, their inclusion could result in beta estimates that depend more on the specific manipulation carried out in each study, which could potentially confound the aggregation of results across different studies. Second, most high-level conditions coded for semantic dimensions, which could represent part of the information that should be captured by semantic models. Hence, including these as covariates could inadvertently remove meaningful signal and hinder model comparison (since semantic models might differ in their sensitivity to the semantic dimensions underlying high-level conditions).

Representational similarity analyses

Contrary to univariate approaches, RSA does not relate neural activity to a *property* of the stimulus or task (first-order isomorphism); rather, RSA is concerned with whether neural activation patterns reflect the representational structure of the stimuli (second-order isomorphism), namely the dissimilarity among stimuli in the space of theoretical models (figure 4.2; Kriegeskorte et al., 2008). The following RSA pipeline was implemented (for the additional details of specific analyses, see section 4.3). First, given a region of interest (ROI), a representational dissimilarity matrix (RDM) was constructed for each subject. These brain RDMs were constructed by computing the Mahalanobis distance between pairs of stimulus phrases, each described by a vector of beta estimates (Nili et al., 2014). From all RDMs, rows and columns corresponding to stimuli with unreliable beta estimates were removed. First, TsDiffAna was used to detect volumes with anomalous activation. If any of these volumes fell within 2 to 10 sec after the onset of a stimulus, the beta estimates for that stimulus were deemed unreliable, and their corresponding row and column was removed from the RDM. Second, single-trial beta estimates of adjacent trials can be unreliable due to collinearity, leading to inflated positive correlations between temporally adjacent trials (Mumford et al., 2014). Because phrases from 1/6 adjacent trials share a constituent in Graves et al., (2010; experiment 1), and phrases from 1/3 adjacent trials share their semantic relation in Wang et al., (2020), the overestimated similarity of adjacent trials might lead to inflated Type I errors (Mumford et al., 2014). Therefore, the first off-diagonal elements were removed from all RDMs. Then, given a ROI and a theoretical model, we quantified their relatedness by computing the Spearman correlation coefficient between their RDMs (Kriegeskorte et al., 2008; Nili et al., 2014). To quantify the relatedness of a ROI and a

theoretical model controlling for the influence of other models, partial Spearman correlations among their RDMs were computed. Finally, I conducted statistical inference over the entire sample of participants and studies by computing right-tailed Wilcoxon signed-rank tests, testing whether the correlation coefficients were significantly greater than zero (Kriegeskorte et al., 2008).

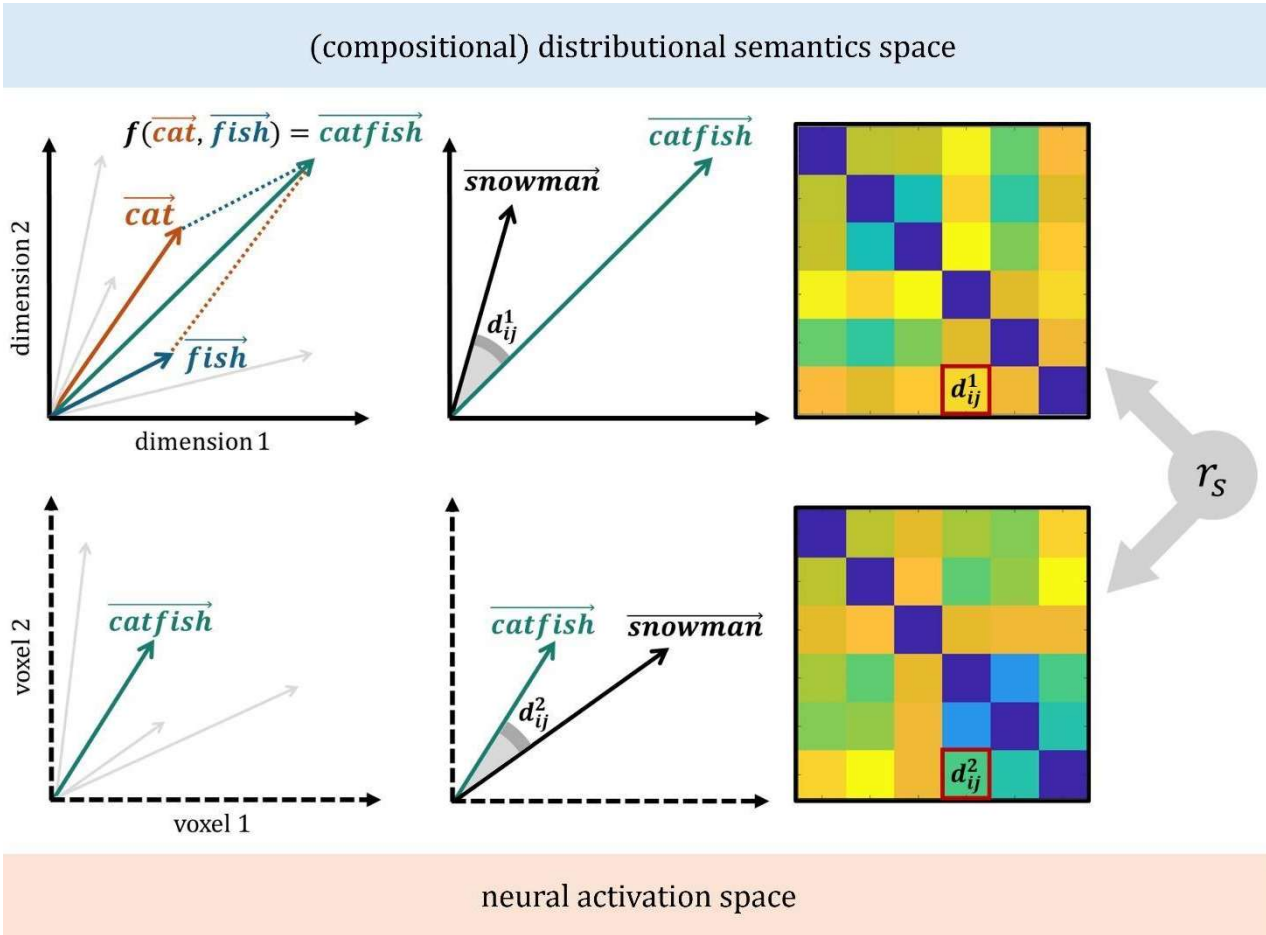

Figure 4.2: Representational similarity analysis (RSA) between the space of a cDSM (top) and the space of a ROI (bottom). The vectors of "cat" and "fish" are combined by a cDSM (the function *f*) into "catfish", whose meaning is compared to another combination, "snowman", via a dissimilarity metric (e.g., $1 -$ cosine similarity), resulting in the dissimilarity value d_{ij}^1. A representational dissimilarity matrix (RDM) is populated by the dissimilarities of all stimuli pairs. The same is repeated in an activation space, where the meaning of combination is defined by the signal of target voxels in response to the presentation of the stimulus. Finally, representational similarity between the cDSM and ROI spaces is quantified by the Spearman rank correlation (r_s) of their RDMs.

Table 4.2: Summary of the characteristics of image acquisition for the four studies included.

study	**Scanning and trial structure**	**image acquisition**
Graves et al., (2010; experiment 1)	The scanning session was split into four runs. On each trial, a phrase was displayed for 1 sec and replaced with a	3T GE Excite system scanner. Functional volumes were acquired using a gradient-echo echoplanar sequence (matrix = 64×64, 33 interleaved axial slices,

	fixation cross. For each run, 120 phrase trials were randomly interspersed with 100 fixation trials (mean duration: 3.6 sec, SD: 2.4).	3.5 mm isotropic voxels, TR = 2000 msec, TE = 25 msec). T1 structural images acquired as a set of 134 contiguous axial slices using a spoiled-gradient-echo sequence. Functional volumes were acquired with gradient-echo echoplanar sequence.
Graves et al., (2010; experiment 2)	Identical to Graves et al., (2010; experiment 1).	Identical to Graves et al., (2010; experiment 1)
Wang et al., (2020)	The scanning session was split into four runs. On each trial, a phrase was shown for 2 sec, followed by a variable duration fixation cross (ranging from 1 to 6 sec).	3T Siemens MRI scanner. Functional volumes acquired using an EPI sequence (matrix = 64×64, 34 descending axial slices, 3.5 mm isotropic voxels, TR = 2000 msec, TE = 22 msec). T1 structural images acquired with gradient-echo inversion pulse sequence (1-mm isotropic voxels).
Graves et al., (2022)	The scanning session was split into four runs. On each trial, a phrase was displayed for 1.5 sec, followed by a fixation cross for 0.5 sec and a fixation cross for a randomly jittered inter-trial interval (mean duration: 2.5 sec).	3T STIM Trio scanner. Functional volumes acquired as T2*-weighted gradient-echo single-shot EPI (matrix = 64 × 64, 35 interleaved axial slices, 3 mm isotropic voxels, TR = 1900 msec, TE = 2.52 msec). T1 structural MPRAGE images (1 mm isotropic voxels).

4.2.3 Theoretical models

Three classes of theoretical models were defined: baseline models coding for possible confounding variables, distributional semantics models coding for non-compositional meaning, and cDSMs coding for semantic composition. Each model defined a theoretical RDM (or candidate RDM; Nili et al., 2014) used in RSA analyses. Figure 4.3 displays the second-order dissimilarities among theoretical models.

Baseline models

To better isolate semantics from confounding representations, five theoretical RDMs were constructed based on the following information:

- **Run:** The run RDM coded for whether trials belonged to the same (0) or different (1) functional runs.

- **Orthographic overlap:** The degree of orthographic overlap between two phrases was quantified by their Levenshtein distance, namely the minimum number of single character edits needed to change a phrase into the other. The orthographic overlap RDM was populated by the Levenhstein distances among all pairs of phrases.
- **Left and right constituent frequency:** Word frequencies were based on the log-transformed word frequency data from the Baroni et al., (2014) corpus, the same corpus used to train the semantic models. I constructed a RDM coding for the absolute difference in word frequency of the first constituents from all trial pairs (e.g., the absolute value of the difference between the word frequency of "snow" and "cat" for the "snow man" and "cat fish" trials). A RDM for the right constituents was constructed analogously.
- **Concreteness:** I employed Brysbaert et al., (2014) concreteness norms and defined the concreteness of a word pair to be the average concreteness of its constituents[13]. A concreteness RDM was constructed based on the absolute difference in concreteness between trials.

Distributional semantics models

Word vectors were obtained from the semantic spaces provided by Baroni, Dinu, et al., (2014); specifically, word vectors were derived from a ~ 2.8 billion word corpus obtained from the concatenation of the web-collected ukWaC (Baroni et al., 2009), a 2008 English Wikipedia dump, and the print-media-based British National Corpus (BNC consortium, 2007). Vectors were computed with the *cbow* version of the *word2vec* model (Mikolov et al., 2013), a prediction-based DSM (Mandera et al., 2017) aimed at predicting a word given its context. The model was trained on the corpus described above, choosing the parameter set producing the best empirical results in Baroni et al., (2014), namely where 400-dimensional vectors were trained to predict words within a context window of five words (negative sampling with k = 10, subsampling with t = 1e−5). These vectors represent the lexicalized meaning of single words, and form the basis on which compositional vectors were generated. Based on these vectors, three semantic models were defined:

- **Left constituents**: $\overrightarrow{w_1}$. These vectors represented the meaning of left constituents, i.e., the first word of each phrase;
- **Right constituents**: $\overrightarrow{w_2}$. These vectors represented the meaning of right constituents, i.e., the second word of each phrase;

[13] Concreteness ratings were missing for 6 first words and 18 second words. In these cases, only the rating of the available constituent was considered.

- **Additive model**: $\overrightarrow{add} = \overrightarrow{w_1} + \overrightarrow{w_2}$ (Mitchell & Lapata, 2010). These vectors represented the unweighted sum of the left and right constituents, or the co-representation of their meanings.

The semantic dissimilarity of two stimuli was quantified by computing 1 – the cosine similarity of their vectors. For each semantic model, a RDM coded for the semantic dissimilarity of stimuli, resulting in three semantic RDMs: w_1 RDM, w_2 RDM, and add RDM.

Compositional distributional semantics models

From the semantic space described in the previous section, two cDSMs were defined:

- **Multiplicative model**: $\overrightarrow{mult} = \overrightarrow{w_1} \odot \overrightarrow{w_2}$ (Mitchell & Lapata, 2010).
- **CAOSS model**: $\overrightarrow{CAOSS} = \overrightarrow{c_1} + \overrightarrow{c_2} = M \times \overrightarrow{w_1} + H \times \overrightarrow{w_2}$ (Marelli et al., 2017). Following Günther & Marelli, (2021), the training data consisted of a set of triplets of (modifier, head, compound) from the union of four noun-noun compound databases: The 629-compound database by (Juhasz et al., 2014), the 1,865-compound database by Günther & Marelli, (2018), the 2,861-compound database by Kim et al., (2019), and the 8,376-compound database by Gagné et al., (2019). To guarantee reliable model representations, I kept those for which the compound and its constituents had a raw frequency higher than 50 in the Baroni et al., (2014) corpus. This resulted in a training set of 4,429 triplets. CAOSS parameters were estimated with the DISSECT toolkit (Dinu et al., 2013a).

$mult$ and $CAOSS$ RDMs coded for the trials' semantic dissimilarity as estimated by these models (i.e., 1 – the cosine similarity of multiplicative and CAOSS representations).

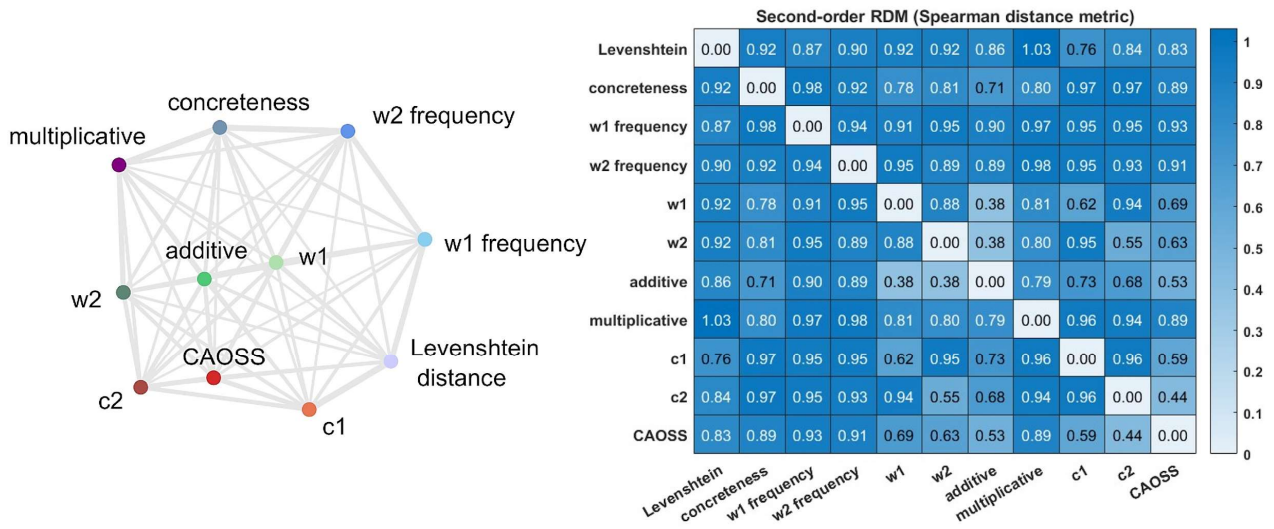

Figure 4.3: Second-order similarities among theoretical models for all stimuli across all studies (run RDM excluded due to subject- and study-specific construction). (left) Multidimensional scaling based on Spearman distance, computed with Nili et al., (2014) toolbox. (right) Second-order dissimilarity matrix; each cell shows the distance of two theoretical RDMs, computed as 1 – their Spearman correlation.

4.3 Analyses and results

4.3.1 Core ROIs

RSA procedure

Complete dataset: RSA was performed in four regions of interest, namely LATL, LAG, RAG, and LIFG. The space of each ROI was defined by a sphere of voxels (diameter = 9 voxels) centered on the mean peak activation from previous fMRI studies. The MNI coordinates used were: LATL (x = -48.1, y = 0.1, z = -24.5), LAG (x = -40.2, y = -66.6, z = 36.9), RAG (x = 46.9, y = -62.3, z = 35.2), and LIFG (x = -47.4, y = 21.3, z = 7.7). The second-order similarity between the representational spaces of a given ROI and theoretical model were quantified by taking the Spearman correlation between their RDMs (Kriegeskorte et al., 2008). In order to control for the influence of confounding dimensions, *partial* RSA were performed, correlating RDM pairs controlling for the influence of other RDMs. Specifically, RSA was performed in two steps:

i) For each ROI, a partial RSA was performed for every candidate semantic (w_1, w_2, add) and compositional semantic ($mult, CAOSS$) model, controlling for baseline models (i.e., including all baseline models as control variables in partial Spearman correlations). Partial Spearman correlation coefficients were then aggregated for all subjects and studies, and a

right-tailed Wilcoxon signed-rank test was performed over these coefficients (Kriegeskorte et al., 2008; Nili et al., 2014), applying Bonferroni correction for the number of ROIs (i.e., $\alpha = .05/4$). Significant results at this stage were taken to indicate that a given ROI represents semantic information consistently with a theoretical model, and that this second-order similarity cannot be attributed to the confounding dimensions of word frequency, orthographic overlap, concreteness, or run. However, at this stage it cannot be known whether the significant results for a compositional model are driven by semantic composition over and above single-word (non-compositional) semantics.

ii) Thus, if either $mult$ or $CAOSS$ were significant in a given ROI, a second partial RSA was conducted controlling for all baseline models and all semantic models. Moreover, if $CAOSS$ was significant, I planned to infer whether this result might be driven by first or second constituents by looking at the results of each, separately (i.e., partial RSA for c_1 and c_2 RDMs).

Non-semantic subset: Because evidence suggests that (compositional) semantic processing is automatically enacted (Günther & Marelli, 2020; Liuzzi et al., 2021), I expected to find a representational match for semantic and compositional semantic models even when composition was not required to carry out a task. To test this hypothesis, I focused on Graves et al., (2010) experiment 1, where participants were tasked to indicate if either word in the current phrase matched a word in the same position from the previous phrase. This 1-back task could be accomplished by simply processing surface-level properties of the stimuli, without semantic processing. Therefore, I considered all semantic and compositional models significant at step i), and repeated the RSA analyses using only this data subset.

Results

Complete dataset: Table 4.3 reports the result of semantic and compositional semantic models (see table A.2.1 for the performance of baseline models). Concerning the former, significant results are observed for the additive model in LAG ($Z = 3.56, N = 85, p < .001$), probably driven, at least in part, by the representational match for the second constituent ($Z = 2.89, N = 85, p = .002$). For compositional models, significant results were observed for the multiplicative model in LIFG ($Z = 2.30, N = 85, p = .011$), and a match close to the adjusted alpha threshold ($\alpha = .0125$) in LATL ($Z = 2.09, N = 85, p = .018$). I then focused on LIFG and on the multiplicative model, and performed a partial RSA controlling for both baseline and semantic models. Results were still significant after this additional control ($Z = 1.76, N = 85, p = .039$), suggesting that the

representational match of the multiplicative model in LIFG is driven by compositional semantics over and above constituent word meaning.

Non-semantic task subset: Focusing on the non-semantic task subset of Graves et al., (2010; experiment 1), I repeated a RSA for the second constituent and additive model in LAG, and for the multiplicative model in LIFG. Neither the second constituent ($Z = 0.99, N = 23, p = .16$) nor the additive model ($Z = 1.14, N = 23, p = .13$) were significant. However, the multiplicative model was still significant in LIFG ($Z = 2.21, N = 23, p = .014$), and remained such after controlling for non-compositional semantic models ($Z = 2.08, N = 23, p = .019$).

Table 4.3: Right-tailed Wilcoxon signed-rank test results for the semantic and compositional semantic models. Test statistics were computed over the partial Spearman correlation coefficients (i.e., subject-level RSA results) for all subjects and all studies (N = 85). All models were controlled for baseline models. The table reports Z scores and uncorrected *p*-values; results significant after Bonferroni correction (adjusted $\alpha = .0125$) are reported in **bold.**

RDM predictor	**LATL**		**LAG**		**RAG**		**LIFG**	
	Z	p	Z	p	Z	p	Z	p
w_1	1.50	.066	1.21	.113	0.51	.306	*2.14*	*.016*
w_2	0.62	.268	**2.89**	**.002**	0.14	.444	1.48	.069
add	1.39	.082	**3.56**	**<.001**	1.32	.094	2.00	.023
$mult$	*2.09*	*.018*	1.32	.093	0.54	.295	**2.30**	**.011**
$CAOSS$	0.21	.418	0.50	.309	0.36	.358	0.02	.493

4.3.2 Searchlight RSA

Semantic network

Procedure: Searchlight RSAs were carried out for spherical ROIs (radius = 3 voxels) centered at each voxel across a target brain region in the space of each subject, for every subject and study. For each sphere, the corresponding brain RDM coded for the Mahalanobis distance between pairs of stimuli. Stimuli falling in outlier volumes were discarded. I performed Spearman correlations between brain RDMs and each baseline model, partial Spearman correlations between brain RDMs and each semantic model controlling for all baseline models, and partial Spearman correlations between brain RDMs and each compositional semantic model controlling for all baseline models and

all semantic models. Searchlights RSAs were implemented with Nili et al., (2014) toolbox. The resulting correlation maps were normalized to reference MNI space, and masked using the intersection of two masks: i) a mask excluding all voxels missing from at least 9 subjects; ii) a mask of the general semantic network peaks from Jackson, (2021) meta-analysis. Group-level analyses over the resulting normalized correlation maps were run using SPM12, applying a threshold of $\alpha = .005$ at the voxel level, and $\alpha = .05$ at the cluster level, family-wise error (FWE) corrected for multiple comparisons.

Results: Across all models, a significant representational match was observed only for the multiplicative model. Specifically, searchlight RSA partialling over both baseline and non-compositional semantic models revealed a significant cluster located in the left middle-superior temporal lobe. When only baseline models were controlled for, a second cluster emerged in *pars triangularis* of the LIFG (BA45). Tables A.2.2 and A.2.3 report the coordinates and test statistics.

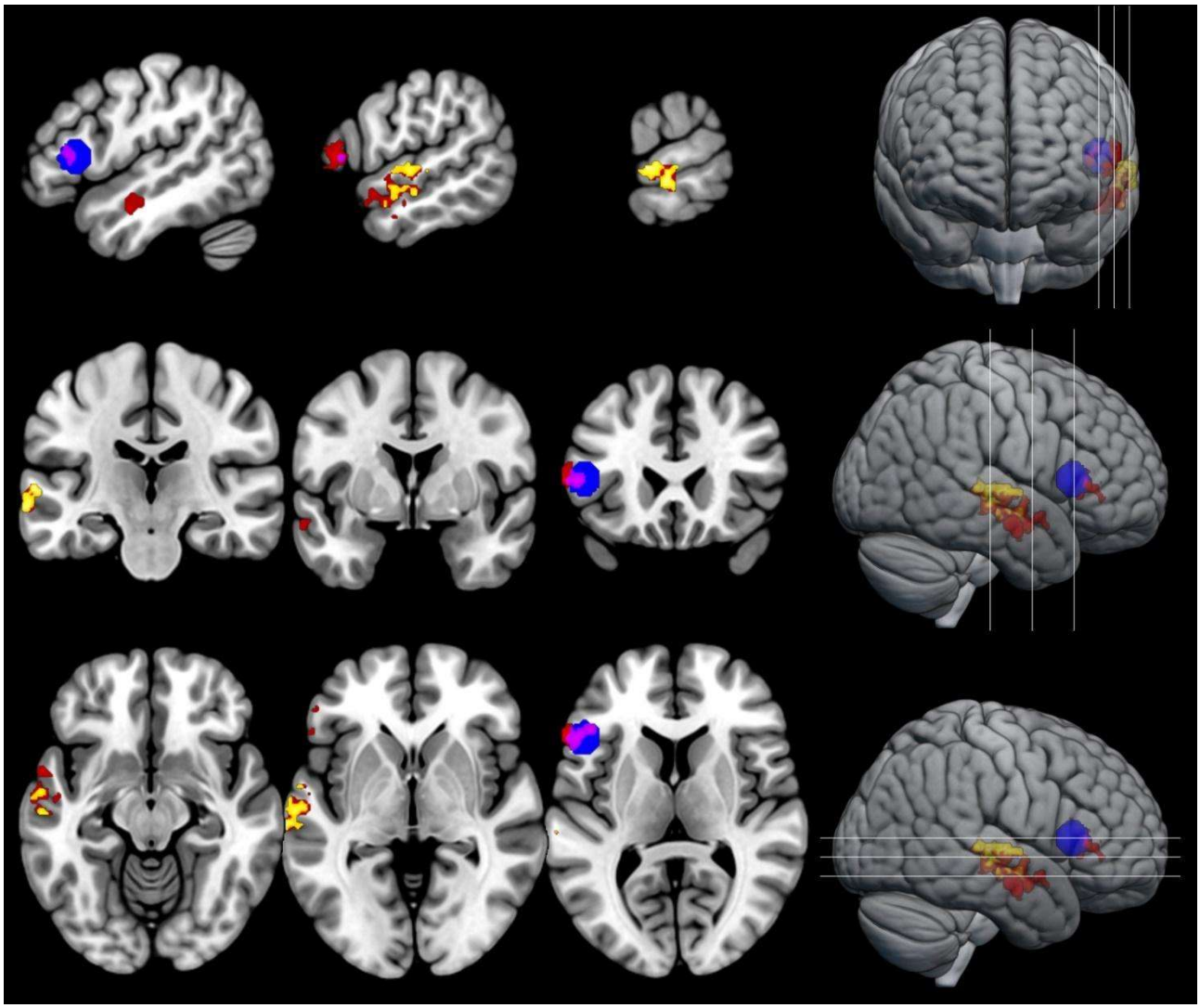

Figure 4.4: Searchlight RSA results in the broad semantic network. Yellow cluster: multiplicative model representations controlling for non-compositional semantic information; red cluster: multiplicative model representations without controlling for non-compositional semantic information; Blue sphere: *a-priori* ROI used in confirmatory RSA.

Semantic network: Attested compounds

Procedure: I repeated the searchlight RSAs in the semantic network focusing on the subset of stimuli for which a lexicalized (i.e., non-compositional) vector was available. Specifically, I identified the stimulus phrases for which a vector representation was available in the Baroni et al., (2014) semantic space: 45 stimuli in Graves et al., (2010; experiment 1 and 2), 11 stimuli in Graves et al., (2022), and 4 stimuli in Wang et al., (2020) met this criterion. Due to the limited number of stimuli, this analysis focused on data from Graves et al., (2010; experiment 1 and 2; N = 45). Brain RDMs coded for the Mahalanobis distance between these 45 stimuli (990 pairs), excluding stimuli falling in outlier volumes. Brain RDMs were constructed from voxel spheres (diameter = 3 voxels) whose center was systematically varied to cover the whole brain.

In this analysis, theoretical models also included the lexicalized representations of phrase stimuli, $\vec{L}$, from which the corresponding L RDM was constructed. Thus, the semantic representations of a given stimulus phrase (e.g., "apple tree") were the following: left constituent ($\overrightarrow{apple}$), right constituent ($\overrightarrow{tree}$), additive model ($\overrightarrow{apple} + \overrightarrow{tree}$), multiplicative model ($\overrightarrow{apple} \odot \overrightarrow{tree}$), CAOSS model ($M \times \overrightarrow{apple} + H \times \overrightarrow{tree}$), and lexicalized meaning ($\overrightarrow{apple\ tree}$). I was especially interested in the unique contribution of the compositional semantic and the lexicalized meaning models. To tease apart the two, partial RSAs were conducted controlling for the confounding dimensions of each model. Thus, given a searchlight voxel sphere, I performed partial Spearman correlations between its brain RDM and the lexicalized meaning RDM, controlling for all baseline, semantic and compositional semantic models. Similarly, I performed partial Spearman correlations between the brain RDM and each compositional model (i.e., $mult$ and $CAOSS$, independently) controlling for baseline models, semantic models, and the lexicalized meaning. The resulting partial correlation maps were normalized to standard MNI space, and masked with a brain mask and a mask excluding all voxels missing from at least 5 subjects. Once again, group-level analyses were performed over the resulting maps applying an alpha threshold of $\alpha = .001$ at the voxel level, and of $\alpha = .05$ at the cluster level, FWE-corrected for multiple comparisons.

Results: No significant clusters were observed.

4.4 Discussion

This chapter explored the neural bases of conceptual combination by testing the representational similarity of combinatorial stimuli described by the activation patterns in core regions of interest and in the space defined by (c)DSMs. To better target semantic representations beyond specific processing demands, fMRI data aggregated from four published studies, all employing two-word combinations but differing in task requirements, was re-analysed. Converging evidence from confirmatory and exploratory RSA (Kriegeskorte et al., 2008) points to a cluster of combinatorial representations of the multiplicative cDSM in left inferior frontal gyrus, suggesting that the latter represents combinatorial information beyond task demands.

Confirmatory analyses: Confirmatory RSA showed that no compositional model was significant in LATL, with the multiplicative model being non-significant after correcting for multiple comparisons ($p = .072$ after Bonferroni correction). This finding is in line with the hypothesis that the LATL specifically supports composition based on the *intersective conjunction* of constituent meaning (Coutanche et al., 2018; Poortman & Pylkkänen, 2016), operationalized by the elementwise multiplication of semantic vectors (Baroni, 2013). While not reaching significance, I contend that this result is still notable in light of the signal dropout in ATL (Visser et al., 2010). Nonetheless, this result must be taken very cautiously, since – besides not reaching significance – exploratory searchlight RSA failed to identify significant clusters in LATL, including those of non-compositional models (going against the vast evidence of LATL centrality in semantic processing). Contrary to my hypotheses, no compositional representations were observed in the angular gyri. The (lack of) finding is consistent with Humphreys et al., (2021) theoretical account of AG, which posits that, rather than representing semantic information *per se*, the AG serves as a temporary buffer of multimodal information. Of course, it is logically possible that AG does support semantic composition, and the cDSMs employed simply did not provide the correct formalism. Turning to the LIFG, significant representational similarities were found between the multiplicative model and this region, which remain significant after controlling for non-compositional semantic models, suggesting that the LIFG is representing genuinely combinatorial information. While combinatorial processing could be confounded with semantic control on specific tasks, leading to spurious representational similarities driven by task demands, the same pattern is less likely to drive similarities across multiple, different tasks. Thus, the fact that multiplicative representations were found in the aggregate dataset suggests that LIFG represents semantic compositions beyond specific task demands. Notably, the searchlight

RSA identifies a significant cluster for the multiplicative model in LIFG[14] which partially overlaps with the ROI mask from the confirmatory RSA analysis in *pars triangularis* (BA45). This localization is consistent with Schell et al., (2017, 2022), who directly contrasted syntactic and semantic phrasal composition. Indeed, the multiplicative model is symmetric (i.e., $\overrightarrow{mult} = \overrightarrow{w_1} \odot \overrightarrow{w_2} = \overrightarrow{w_2} \odot \overrightarrow{w_1}$), and hence characterizes combinations semantically, not syntactically. Finally, multiplicative representations – controlling for non-compositional semantic – were observed even in the 1-back data subset from Graves et al., (2010; experiment 1), suggesting that combinatorial processing is automatically enacted by LIFG even when semantic access is not required to carry out the task at hand, consistent with Liuzzi et al., (2021) findings of automatic semantic representation in LIFG *pars triangularis*.

Exploratory analyses: Searchlight RSA identified significant clusters of multiplicative representations in BA45 and the left middle-superior temporal lobe; only the latter remain significant after controlling for non-compositional semantic information. Extensive evidence links both regions to semantic memory (Binder et al., 2009; Jackson, 2021; Ralph et al., 2016), with both regions shown to represent semantics as modeled by DSMs (Carota et al., 2017; Liuzzi et al., 2020; Zhang et al., 2020). Looking more specifically at composition, however, previous works implicate more posterior portion of the left temporal cortex (Baggio, 2018; Graessner et al., 2021; Molinaro et al., 2015), while evidence for the involvement of the middle-superior portion points to role binding (i.e., "who did what to whom"; Frankland & Greene, 2020a, 2020b), which is incompatible with the multiplicative model's symmetry. Thus, these results are more challenging to reconcile with existing evidence for the localization of semantic composition. However, the latter are overwhelmingly based on univariate analyses, and a degree of divergence between the former and the latter is, in this sense, expected.

Modeling considerations: It is notable that, across both confirmatory and exploratory analyses, the multiplicative model consistently emerged as the only cDSM and, in the exploratory RSA, the only theoretical model displaying significant representational similarities with target brain regions. Because cDSMs were tested against data coming from diverse study settings, which might share only the most fundamental properties, I speculate the good performance of the multiplicative model might stem from its simplicity. Nonetheless, while also simple, representational similarities with the additive model were not observed as pervasively and consistently as those of the multiplicative model. Therefore, at least compared to addition, the superior performance of the

[14] This cluster emerges only when the multiplicative representations are not controlled by non-compositional semantic models, suggesting the latter do share relevant variability with the multiplicative model (figure 4.3), despite not being significant when taken alone.

multiplicative model cannot be reduced to simplicity. Instead, the CAOSS model was never significant, despite predicting compound processing and representation across different behavioral measures and languages (Günther, Marelli, et al., 2020; Günther & Marelli, 2018, 2020, 2022; Marelli et al., 2017; Chapter 2 and 3). Mirroring the considerations of model complexity advanced above, the CAOSS model might be too complex to generate semantic representations that are consistently present across tasks, and thus preserved in their aggregate. Alternatively, while CAOSS predicts a diverse array of behavioral results related to compound processing, the representations that drive such predictions may just not be good models of the neural representations that underlie those behaviors, suggesting that more is needed to develop a cDSM that can account for both.

Limitations and conclusions: Considering data coming from diverse fMRI studies made it possible to conduct analyses at a scale larger than usual experiments, posing a stringent test to the generalizability of computational representations across contexts. This advantage came at the cost of data quality, as the studies re-analyzed were not optimized for MVPA, leading to noisy beta estimates and suboptimal BOLD acquisition in LATL. Studies tailored to the present research questions and methodology would have, of course, more Power to detect the effects of interest, a limitation that could have partially been mitigated by the large sample size. In this sense, the reduced sample size (45 subjects from Graves et al., (2010); experiment 1 and 2) might be the reason for the failure to observe significant clusters in the searchlight analysis for attested compounds.

Finally, discussed in the introduction, the present study attempted to shift the focus of neuroimaging studies of semantic composition from an "activation" perspective, relating univariate responses to high-level contrast that involve combination, to a "representational" one, where semantic composition is defined as a systematic transformation of meaning-carrying representations. Three key ingredients characterize this framework: the representations of constituent concepts, the representation of their combined meaning, and - the end goal of the research endeavor - the transformation that connects the two. Reverberi et al., (2012) provide an example of this approach in the domain of rule representation. Specifically, they obtained the multivoxel BOLD representations of conditional rules and their conjunctive combination, showing that decoders trained on constituent rule identity could be used to successfully recover the identity of their combination. From this perspective, an ingredient missing from the present study is the neural representation of constituent information. Future studies might collect the neural representation of constituent units besides those of their combination; cDSMs could then directly operate on the former to predict the latter (indeed, see Baron et al., 2010; Baron & Osherson, 2011).

Chapter 5

Compounding in large language models

5.1 Introduction

The CARIN (Gagné & Shoben, 1997) and RICE (Spalding, Gagné, Mullaly & Ji, 2010) theories propose that conceptual combination is supported by a set of implicit semantic relations that compete during processing to determine a relational interpretation (e.g., "man MADE OF snow"). These theories posit that the adequacy of an interpretation depends on its ability to conform to a system of *conceptual roles* (or concept affordances; Lynott & Connell, 2010) involving both constituents, their interaction, and their relations with the broader knowledge base (Gagné & Spalding, 2013; Lake & Murphy, 2021; Maguire et al., 2010). For example, "snow" can be a modifier argument of the MADE OF relation because it is a material, while it likely cannot function as a modifier of the USED BY relation, since it is not an agent. However, in the case of "snowboard", "snow" is less likely to be an argument of MADE OF, since boards require firmness. Whether semantic relations form the basis of all combinations, as CARIN and RICE theories claim, or only of some (e.g., "relational" combinations in the dual-process model; Wisniewski, 1996), most theories hold that semantic relations are essential to characterize conceptual combination (section 1.2.1). Consequently, computational models of conceptual combination should be able to account for implicit relational information.

In this context, cDSMs appear limiting. Indeed, cDSMs define combination via arithmetic or linear-algebraic operation over constituent vectors and, while these can be generally interpreted based

---

This work was presented at the 2023 edition of the "Psycholinguistics in Flanders" conference (University of Ghent, Ghent, Belgium) as the following oral contribution: "Ciapparelli, M., & Marelli, M. (2023). Modeling compound word relational interpretations with contextualized word embeddings".

This work was also presented at the 2023 "Conference of the European Society for Cognitive Psychology" (ESCOP; University of Porto, Porto, Portugal) as the following oral contribution: "Ciapparelli, M., Zarbo, C., & Marelli, M. (2023). Contextualized word embeddings capture compound words' implicit relational interpretations."

This work was presented at the 2024 "International Word Processing Conference" (Belgrade University, Belgrade, Serbia) as the following oral contribution: "Ciapparelli, M., Zarbo, C., & Marelli, M. (2024). Uncovering implicit relational interpretations of compound words with contextualized word embeddings."

on how they combine vector dimensions (see Chapter 4), they do not directly correspond to specific relations. In other words, there is no mathematical *operation* in semantic space that captures the MADE OF relation and hence no model that can generate the meaning of "[head] MADE OF [modifier]" (although relations can be defined as directions in semantic space; Levy & Goldberg, 2014). While it is possible to train tensors (matrices) for each relation (i.e., as in the example of adjectives; section 1.3.2), this would require training a model for each relation, a solution that would not be flexible (i.e., immediately adaptable to new semantic relations) nor parsimonious. Instead, a single model should ideally be able to represent a compound as well as the interpretations that result from applying different semantic relations to connect its constituents. In this context, large language models (LLMs) present a compelling modeling alternative. LLMs take a top-down approach privileging the representation of sentences (Pavlick, 2022), which are broken down into token components connected by attention heads (Vaswani et al., 2017). This framework allows LLMs to generate good representations for longer sequences of text and, in principle, blur the distinction between the meaning of words, complex words, and their interpretations (e.g., "snow", "snowman", and "a man made of snow"; see general introduction 1.3.3). Moreover, while LLMs have been criticized for their lack of direct grounding (Bender & Koller, 2020; Lake & Murphy, 2021), their ability to model the relationship between concepts, or "conceptual roles", is less disputed (Piantadosi & Hill, 2022). On this account, given that part of what constitutes word meaning in general, and combinatorial meaning in particular, is relational, and since many of the tasks that LLMs succeed on are those pertaining to relational information (Piantadosi & Hill, 2022), then LLMs should be expected to account for compound words' relational interpretations.

Nonetheless, there are also reasons to question the ability of LLMs to capture finer grained relationships among concepts. Indeed, Lake & Murphy, (2021) contend that LLMs characterization of semantic relations is simplistic, since they resort on general metrics of similarity (e.g., cosine similarity), glossing over specific sources of relatedness (e.g., while "cat" may be semantically similar to "whiskers", "feline" and "mouse", the source of these similarities is different, as cats HAVE whiskers, ARE felines and PREY ON mice). Tapping into this level of specificity should be essential to account for compound word meaning (i.e., simply knowing the similarity of "snow" and "man" is not enough to infer the MADE OF relation). Indeed, much of the evidence for the CARIN and RICE theories draws from a taxonomy of semantic relations (see table 5.1; Gagné, 2001; Gagné & Shoben, 1997; Gagné & Spalding, 2009; Spalding & Gagné, 2014), with some studies relying on the *possible relation task*, whereby participants are tasked with selecting the most appropriate semantic relation to interpret a compound. Thus, compounding represents an ideal domain in which LLMs' ability to

represent combinatorial and relational information can be tested at once, and the possible relation task an ideal testing ground.

This chapter is devoted to test LLMs ability to model compound meaning relationally, that is, in terms of a set of possible and implicit meanings and their relative strength, determined by their ability to take on appropriate conceptual roles (Piantadosi & Hill, 2022). To do so, models are tested against data from the possible relation task. First, following the possible relation task setting, the implicit relational interpretation of a contextualized compound is made explicit, e.g., "Yesterday we built a beautiful **man made of snow** in our garden". Here, LLMs' ability to generate representations for longer sequences of text is leveraged to generate contextualized word embeddings (CWE) of the target interpretation. Because we know that the distance between CWEs in a semantic space is inversely proportional to their semantic similarity, the distance between the CWEs of "snowman" in its original and paraphrased forms is considered as an estimate of the adequacy or strength of the paraphrase. If the model is sensitive to the conceptual roles taken on by the compound and its constituents, this distance will depend on the number and extent of conceptual role violations brought by the relation. The procedure is repeated for all relational interpretations, and their model-predicted strength is used to predict the number of times these were selected by human raters. Thanks to the variety of experimental settings provided by Benjamin & Schmidtke, (2023), Schmidtke, Gagné, Kuperman, & Spalding, (2018) and Schmidtke, Gagné, Kuperman, & Tucker, (2018), the procedure is repeated for existing compounds rated in isolation and in context, and for novel compounds rated in context.

5.2 Methods

5.2.1 Large language models

I focused on two LLMs, namely BERT-base (Devlin et al., 2019) and Llama-2-13b (Touvron et al., 2023).

BERT-base: BERT has shown promising results in capturing some of the semantic properties of compounds (semantic transparency and headedness, Buijtelaar & Pezzelle, 2023; literality, Miletic & Schulte, 2023) and, more generally, of words (Cassani et al., 2023). I employed BERT-base-uncased (Devlin et al., 2019) using HuggingFace (Wolf et al., 2020) implementation[15]. BERT-base is

[15] https://huggingface.co/bert-base-uncased

a 110 million parameters Transformer-based model pre-trained on English using the Masked Language Modeling (i.e., predicting words masked in the input) and Next Sentence Prediction (i.e., predicting whether two sentences follow each other) objectives. The model comprises a sequence of 12 encoder blocks, each endowed with 12 attention heads. At each layer, the model learns 768-dimensional CWE. Starting from input token embeddings, attention heads operate in parallel to route information among tokens, resulting in embeddings that are progressively more intertwined with the broader sentential context.

Llama-2-13b: While BERT-base has been extensively shown to encode relevant aspects of word meaning and, at least partially, of combinatorial semantics, more recent LLMs have outperformed BERT-base across virtually all benchmarks (Greco et al., 2022). Moreover, the vast increase in model size seen in recent years resulted in substantial improvements across various tasks, sometimes leading to abrupt or "emergent" improvements (Srivastava et al., n.d.; Wei et al., 2022). For these reasons, I employed the recently released Llama-2-13b (Touvron et al., 2023; HuggingFace implementation[16] at 4-bit quantization[17]), a 40 layers pre-trained 13 billion parameter Transformer-based language model developed to reduce costs at inference by training the model on more data (i.e., 2 trillion tokens for Llama-2). Despite being smaller compared to many LLMs, Llama-2-13b matches the performance of state-of-the-art open-source LLMs on academic benchmarks and, while lagging behind best-performing closed-source LLMs (e.g., PaLM-2-L, GPT-4), matches or outperforms much larger models (e.g., GPT-3; Touvron et al., 2023).

5.2.2 Large language models: Modeling and predicting relational interpretations

The following procedure was adopted across all studies to quantify the probability of selecting a specific relational interpretation for a given compound.

Sentence selection and transformation: To obtain the representation of a target compound, a sentence in which the compound was embedded (see section 5.3 for analysis-specific details) was identified. Then, 16 sentence variants were generated, all identical to the original except for the compound word, which was replaced with a specific relational interpretation. For example, given a source sentence such as: "*Yesterday we built a beautiful* ***snowman*** *in our garden"*, 16 sentences of the type: "*Yesterday we built a beautiful* ***man made of snow*** *in our garden"* were obtained. These 16 interpretations (table 5.1) were drawn from the set of relations from the possible relation tasks in

[16] https://huggingface.co/meta-llama/Llama-2-13b-hf
[17] https://huggingface.co/docs/transformers/main_classes/quantization

Schmidtke, Gagné, et al., (2018), Schmidtke, Van Dyke, et al., (2018) and Benjamin & Schmidtke, (2023). Because some of these relations might be problematic when embedded in a sentential context (e.g., "has") and because alternative wording can affect models' semantic knowledge (Lake & Murphy, 2021), 20 alternative wordings were considered for these relations (table 5.1). Relation probabilities were computed from the cosine similarity values for all resulting 36 relational interpretations. The assumption was that, once substituted for the compound, relations that would result in ungrammatical sentences and less meaningful interpretations would also deviate more from the original sentence compared to fitter alternatives. Thus, when multiple variants of the same relation were available (e.g., "has", "that has" and "having"), the one with the largest cosine similarity was selected.

Table 5.1: Relational interpretations employed. The left column reports the 16 relations originally employed in all possible relation tasks studies. The right column reports the alternative wording for these relations.

Original relational interpretation	Variants			
H about M				
H by M				
H caused by M				
H causes M	H causing M	H that causes M		
H during M				
H for M				
H from M				
H has M	H having M	H that has M		
M has H	M having H	M that has H		
H is M	H being M	H that is M		
H location is M	H whose location is M	H located on M	H located in M	H located at M
M location is H	M whose location is H	M located on H	M located in H	M located at H
H made of M				
H makes M	H making M	H that makes M		

H used by M		
H uses M	H using M	H that uses M

Contextualized word embeddings generation: For each sentence and each variant, the minicons library was used (Misra, 2022) to extract its CWE. Specifically, given a target sequence within a sentence (i.e., a compound or its paraphrase), the minicons CWE module was used to extract its CWE, namely the 768-dimensional vector from a given BERT-base layer and the 5120-dimensional vector from Llama-2-13b. When the target sequence comprised multiple tokens, their embeddings were averaged to obtain the final CWE (Bommasani et al., 2020).

Relation probabilities computation: The main hypothesis was that a more plausible interpretation (*Yesterday we built a beautiful* ***man made of snow*** *in our garden*) would result in a CWE closer to the original sentence (*Yesterday we built a beautiful* ***snowman*** *in our garden*) compared to less meaningful alternatives (*Yesterday we built a beautiful* ***man for snow*** *in our garden*). Thus, the cosine similarity between the original compound CWE and the CWE of each of its 16 variants was computed. Because the main concern was about the *relative* dominance of possible relational interpretations, for each compound:

1. The minimum cosine similarity value was subtracted from the similarity values of all 16 variants. This step emphasizes the *relative strength* of each relational interpretation, neglecting differences among compounds that are irrelevant to the possible relation task.
2. Probabilities over these truncated cosines were then obtained by dividing each value by the sum over the 16 variants after truncation.

$$p(r_k)_i = \frac{\cos(\vec{c_i}, \vec{r_k}) - \min_j(\cos(\vec{c_i}, \vec{r_j}))}{\sum_j^{16}(\cos(\vec{c_i}, \vec{r_k}) - \min_j(\cos(\vec{c_i}, \vec{r_j})))}$$

Where $p(r_k)_i$ is the model-estimated probability of selecting relation r_k to interpret compound i, $\vec{c_i}$ is the CWE of compound word i, and $\vec{r_k}$ is the CWE of compound i paraphrased with relation k from a set of j relations. From now on, these $p(r_k)$ values will be called *relation probabilities.* Figure 5.1 illustrates the procedure in the simplified case of four relational interpretations.

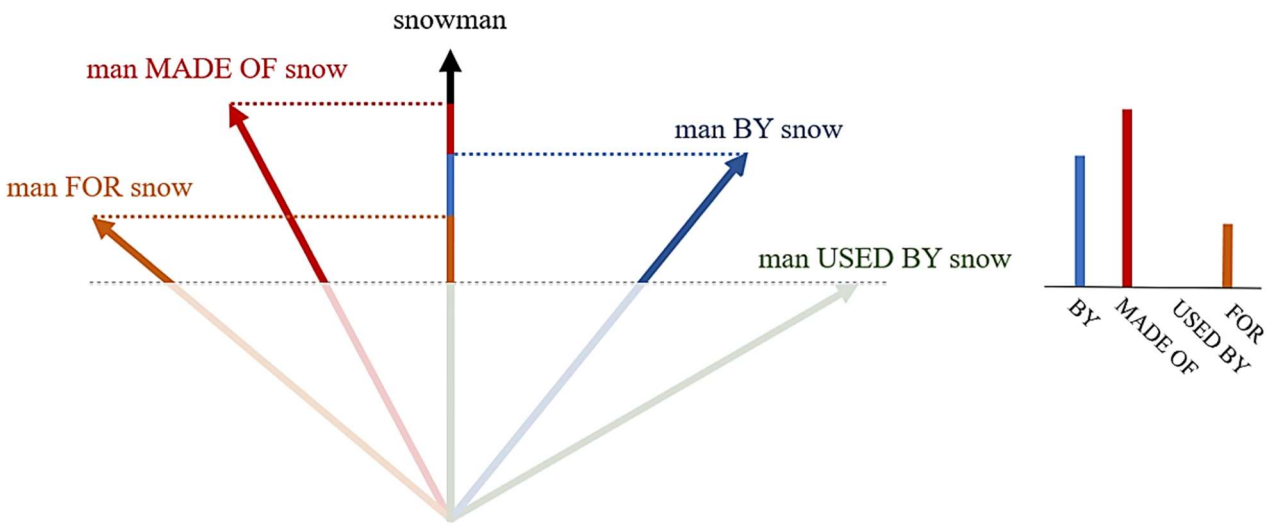

Figure 5.1: Relation probability computation in the simplified case of four relational interpretations. (*left*) The CWE of the compound "snowman" (black vector) is compared to each relational interpretation with cosine similarity (i.e., normalized dot product, or the magnitude of vector projection after vector length normalization). The resulting similarities (colored projections onto the snowman vector) are truncated at the level of the minimum similarity (here, man USED BY snow). (*right*) The truncated similarities are transformed into probabilities (i.e., each truncated similarity is divided by the sum of truncated similarities).

5.2.3 Baseline models

To test whether LLMs improve on previous models, they were compared to baseline models from the compositional distributional semantics tradition. Specifically, the semantic space provided by Baroni, Dinu, et al., (2014) was used, namely a ~ 2.8 billion word corpus obtained from the concatenation of the web-collected UkWaC (Baroni et al., 2009), a 2008 English Wikipedia dump, and the print-media-based British National Corpus (BNC consortium, 2007). Static word embeddings were derived with the *cbow* version of the *word2vec* model (Mikolov, Chen, et al., 2013), choosing the parameter set producing the best empirical results in Baroni et al., (2014), namely where 400-dimensional vectors were trained to predict words within a context window of five words (negative sampling with k = 10, subsampling with t = 1e−5). Using these static word embeddings, three baseline models were defined:

Vector difference: Semantic relations were defined as directions in semantic space computed by vector difference. This approach has been proven capable of capturing semantic relations (Zhang et al., 2020), human judgments of object properties (Grand et al., 2022), and analogical operations (Mikolov, Yih, et al., 2013). Specifically, for a given relational interpretation (e.g., ABOUT), 100

unique common noun pairs[18] interspersed with the target relation in UkWac sentences (e.g., "news ABOUT unemployment") were selected. The embedding of a semantic relation was defined by the difference between the embedding of the second and first noun ($\overrightarrow{about} = \overrightarrow{unemployment} - \overrightarrow{news}$), and averaged the results across all 100 word pair instances. While all relations were specified as "head RELATION modifier", the HAS and LOCATION IS relations included the reverse direction as well (i.e., "modifier HAS/LOCATION IS head"); in these cases, the relation vector's sense was reversed. Then, compound word embeddings ($\vec{c}$) were described by the difference between the embedding of the modifier ($\vec{m}$) and head ($\vec{h}$) constituents:

$$\vec{c} = \vec{m} - \vec{h}$$

Additive model (Mitchell & Lapata, 2010): Semantic relations were defined by the corresponding vector in the Baroni et al., (2014) space (e.g., $\overrightarrow{about}$). For relations composed of two words (e.g., CAUSED BY), the relation's vector was obtained by averaging the vector of each word. Compound word vectors were obtained by the unweighted sum of their modifier and head constituents vectors:

$$\vec{c} = \vec{m} + \vec{h}$$

Compounding as Abstract Operations in Semantic Space (CAOSS) model (Marelli et al., 2017): As for the additive model, semantic relations were defined by their vector in the Baroni et al., (2014) space. Compound word vectors were obtained as follows:

$$\vec{c} = M\vec{m} + H\vec{h}$$

Where the M and H are matrices estimated by a training procedure aimed at predicting compound word vectors from their respective constituents vectors. The CAOSS model has been shown to capture various lexical and semantic features of compound word representation and processing (Günther & Marelli, 2021; Marelli et al., 2017), and to implicitly encode information of the appropriate relational interpretations (Günther & Marelli, 2022). The model was trained on a set of ($\vec{m}, \vec{h}, \vec{c}$) ordered triplets from the union of the English compound word databases of Gagné et al., (2019), Günther & Marelli, (2018), Juhasz et al., (2014) and Kim et al., (2019), restricted to the 4,429 instances for which word vectors had a frequency higher than 50 in the Baroni et al., (2014) corpus. The M and H matrices were estimated with the DISSECT toolkit (Dinu et al., 2013).

[18] The variant "located in" was used for the relation LOCATION IS, for which the greatest number of valid UkWac sentences (i.e., 64) was available.

5.2.4 Baseline models: Modeling and predicting relational interpretations

To estimate relation probabilities with baseline models, I followed the procedure used for LLMs. Specifically, for all baseline models, relation probabilities were quantified by computing the cosine similarity between the embeddings of each compound and semantic relation and transforming the result into probabilities after truncating at the minimum cosine similarity value.

5.3 Models evaluation

5.3.1 Analysis 1: Existing compounds rated in isolation

Dependent variables: The dependent variable of analysis 1 was the number of participants selecting a given interpretation for a given compound in the possible relation tasks from Schmidtke, Gagné, Kuperman, & Spalding, (2018) and Schmidtke, Gagné, Kuperman, & Tucker, (2018). On each trial of the task, a compound word was presented in a spaced format (e.g., "snow man") in isolation. Participants were instructed to pretend they were learning English and were familiar with each of the constituents but never saw them used together. Participants were asked to select the most likely literal meaning for the combination of constituents, choosing one of 16 possible relational interpretations. Schmidtke, Gagné, Kuperman, & Tucker, (2018) collected judgments from 263 raters for 604 attested English compounds (44 median judgments per compound). Schmidtke, Gagné, Kuperman, & Spalding, (2018) collected judgments from 186 raters for another 600 attested English compounds (30 median judgments per compound). I aggregated data from both studies, resulting in a final set of possible relation task judgments for 1204 English compound words.

Model predictions: Bommasani et al., (2020) show that the best strategy to obtain decontextualized representations is to average the CWE of the target word embedded in multiple sentential contexts. Thus, decontextualized embeddings were obtained by averaging CWEs obtained across many naturally occurring sentences. The UkWac corpus (Baroni et al., 2009) was used to search for sentences in which the compounds were embedded. I sorted sentences based on their GDEX score (i.e., an index developed to detect good dictionary examples from a corpus; Kilgarriff et al., 2007) and selected the first 15 sentences that met the following criteria: (i) the compound word was present only once, (ii) the compound word was present in singular and lowercase form, (iii) sentence length could not exceed 200 characters. I did not consider compounds for which less than 15 sentences were available. 1120 out of the 1204 original compounds met these selection criteria.

Therefore, the final dataset comprised 15 sentences for 1120 compounds, each paraphrased in 37 variants (i.e., the original sentence and its 36 alterations, see table 5.1), resulting in 621600 sentences. I considered the CWE of a given compound and its variants to be the average CWE across the 15 sentences[19] in which it occurred and computed relation probabilities following the procedure described in section 5.2.2.

Statistical analyses: The distribution of the count dependent variable follows a decaying exponential profile which we modelled with a negative Binomial link function (Green, 2021). Generalized linear mixed models were fitted using the lm4 package (Bates et al., 2015) for the R software (R Core Team, 2022). Conformance to model assumptions was assessed with the DHARMa package (Hartig, 2020). Relation probability values were standardized and entered as main predictors. Since the median number of participants differed between the two studies, I also included the main effect of study source as a dummy variable, coding for whether ratings were obtained from Schmidtke, Gagné, Kuperman, & Tucker, (2018) or from Schmidtke, Gagné, et al., (2018). Random intercepts for relational interpretation were added to account for baseline differences in their selection frequency. I fitted an independent mixed effects model for each candidate semantic model, namely the three baseline models, the 12 layers of BERT-base, the 40 layers of Llama-2-13b, as well as their control variants (i.e., alternative wording of semantic relations and number of sentences over which the CWEs were averaged). Bonferroni correction was applied for the number of layers of BERT-base ($\alpha/12$ at $\alpha = .05$) and Llama-2-13b ($\alpha/40$). The Akaike information criterion (AIC) was used to perform model comparison. Finally, to control for the impact of overly influential outliers, I performed model criticism on the best-fitting layer of BERT-base and Llama-2-13b, as per AIC. Therefore, I removed residuals greater than |2.5| standard deviations and re-fitted the generalized linear mixed models on the truncated dataset (Baayen, 2008).

Results: The main effect of relation probability survives Bonferroni correction for all BERT-base layers. BERT's performance increases almost monotonically, culminating in the best results observed for the 12th layer, greatly outperforming Llama-2-13b (lowest $\Delta AIC = 645.8$) and baseline models (lowest $\Delta AIC = 1061.1$); see figure 5.2. In line with the main hypothesis, relation probability and selection frequency are positively related. Focusing on the last layer results, figure 5.3 shows

[19] Because Miletic & Schulte, (2023) show that the number of sentences considered for averaging affects BERT's ability to predict compound compositionality, I investigate the extent to which model performance depends on the number of sentences over which the average CWE is computed. Specifically, I repeated the procedure by averaging the first 10, 5, and 2 sentences. For both BERT-base and Llama-2-13b, only a slight decrease in performance is observed, still culminating at layer 12 and 14, respectively (figures A.3.1 and A.3.2).

how frequency of selection varies as a function of relation probability ($\beta = 0.43, p < .0001$; i.e., for every standard deviation increase in relation probability, the number of participants selecting a relational interpretation increases by 53%). Results remain consistent after model criticism ($\beta = 0.46, p < .0001$) and are robust to alternative wording of semantic relations ($\beta = 0.46, p < .0001$). Compared to BERT-base, Llama-2-13b's performance shows an inverse U-shape pattern, peaking at the 14th layer ($\beta = 0.37, p < .0001$). Similarly to BERT, Llama performance is robust to model criticism ($\beta = 0.38, p < .0001$) and to the alternative formulation of semantic relations ($\beta = 0.39, p < .0001$).

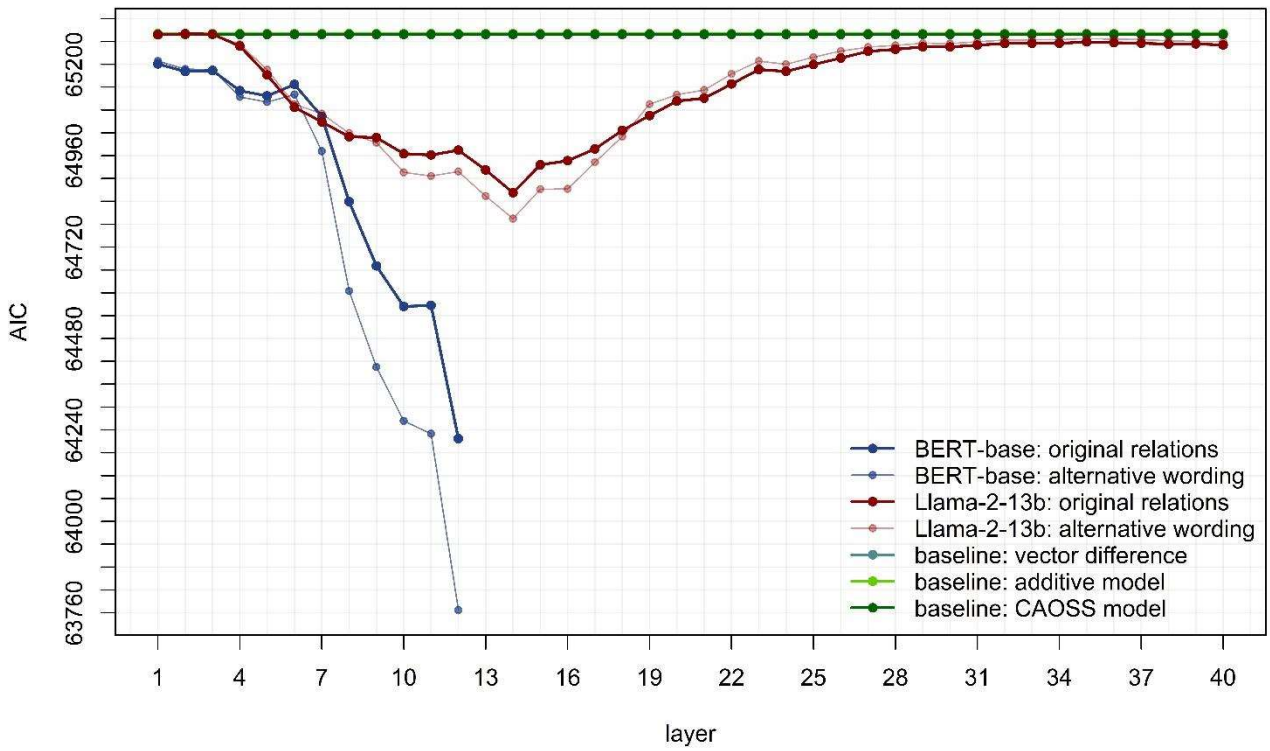

Figure 5.2: AIC values of the generalized linear mixed models with relational probability as main predictor.

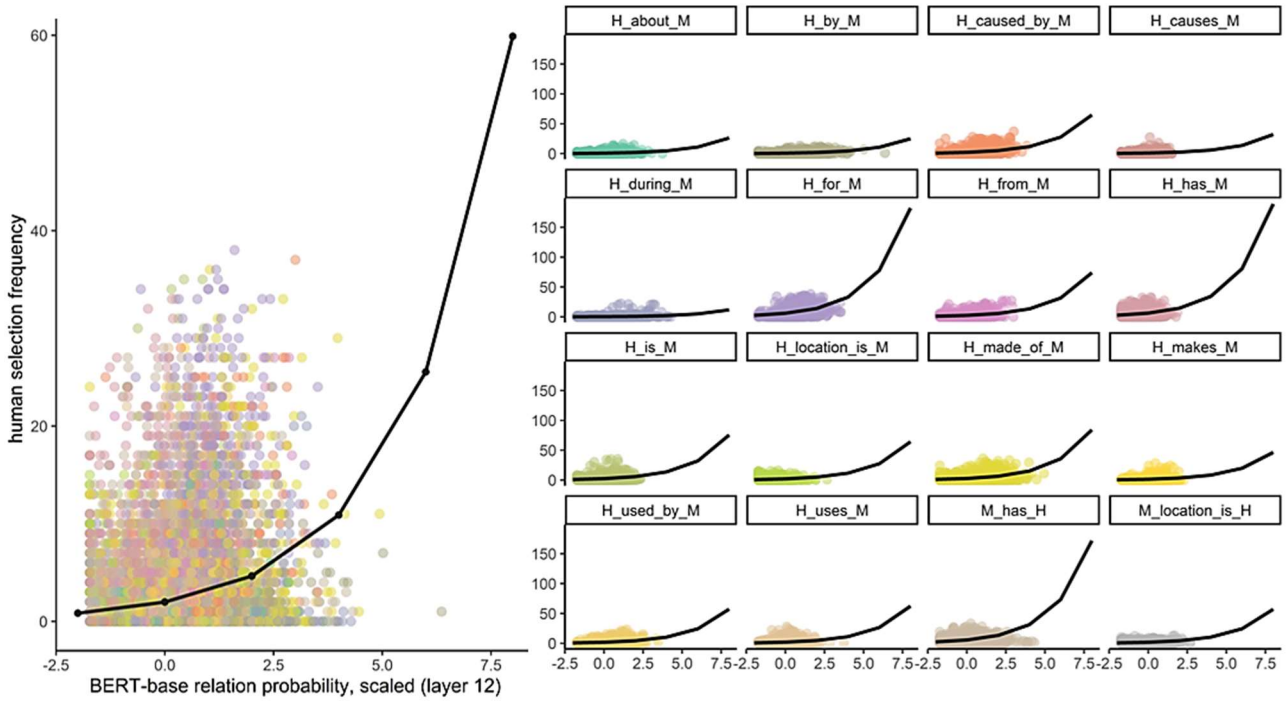

Figure 5.3: Relation probability effects for BERT-base layer 12. (*left*) Main effect of relation probability (scaled) on selection frequency in the possible relation task. Attested compound words were in isolation. (*right*) Main effect decomposed across all 16 random intercept effects for relation.

Interim discussion: In line with predictions, the relation probability of an interpretation is positively related to its selection frequency. The plausibility of a specific interpretation depends on whether it is consistent with a system of conceptual roles involving the constituents, their interaction, and the broader lexical system (Gagné & Spalding, 2013; Maguire et al., 2010). On the one hand, participants pick up on these (in)consistencies to carry out the possible relation task. On the other hand, if LLMs pick up on dimensions or conceptual roles that are also relevant to human judgments, we expect paraphrasing to alter CWE to a degree that is consistent with human ratings. Thus, these results can be taken as indirect indication that LLMs and humans are sensitive to similar underlying dimensions. Of course, it is also possible that the positive relation observed is driven by other spurious factors, which might emerge when the experimental conditions vary. Thus, I then tested whether this finding generalizes to a different experimental condition, and considered a smaller set of possible relation task judgements obtained after each compound was read in a sentential context, which led to more homogeneous interpretations (i.e., lower Shannon entropy; Benjamin & Schmidtke, 2023).

5.3.2 Analysis 2: Existing compounds rated in context

Dependent variables: The dependent variable was the number of participants selecting a given interpretation for a given compound in Benjamin & Schmidtke, (2023). On each trial, a sentence with an embedded compound appeared one word at a time in the center of a screen, with participants progressing to the next word by pressing the space key. After reading each sentence, participants performed a possible relations task for the embedded compound, analogously to the possible relations task for compounds presented in isolation. 55 participants rated 100 existing English compounds, each embedded in two sentences each (26 median judgements per compound and sentence).

Model predictions: The dataset comprised 2 sentences for 100 compounds, each paraphrased in 37 variants, resulting in 7400 sentences. Because compounds were presented in context, I generated BERT-base and Llama-2-13b CWE for the very sentences presented to participants. Relation probabilities were computed for all layers following the procedure described in section 5.2.2.

Statistical considerations: The statistical procedure was identical to that described in section 5.3.1, except for study source: since all data was obtained from Benjamin & Schmidtke, (2023), I did not need to differentiate the source of data with a dedicated covariate.

Results: Again, relation probability and selection frequency were found to be positively related. As in analysis 1, BERT's fit increased almost monotonically, reaching its best performance at the 12th layer, whereas Llama-2-13b performance shows a reverse U-shaped progression, peaking at layer 14 (figures A.3.3 and A.3.4). The main effect of relation probability survived Bonferroni correction across all BERT layers, while only Llama's layers 5 to 29 survive the correction, and only layers 14, 15, and 16 improve the fit of the best-performing baseline model. Overall, BERT's layer 12 achieves the lowest AIC, vastly outperforming all baseline models (lowest $\Delta AIC = 89.2$) and all Llama-2-13b layers (lowest $\Delta AIC = 74.9$). Figure 5.4 shows how selection frequency varies as a function of relation probability for BERT layer 12 ($\beta = 0.56, p < .0001$). Results remain consistent after model criticism ($\beta = 0.60, p < .0001$), and are robust to alternative wording ($\beta = 0.58, p < .0001$; figures A.3.3 and A.3.4).

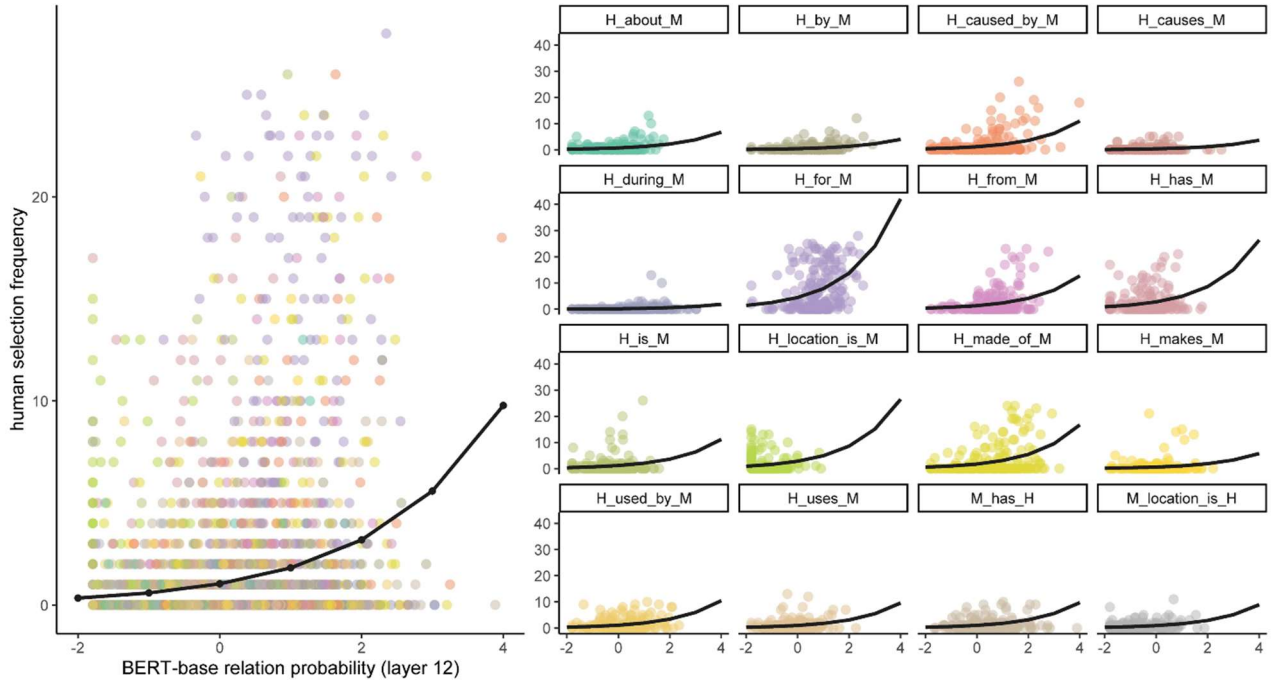

Figure 5.4: Relation probability effects for BERT-base layer 12. (*left*) Main effect of relation probability (scaled) on selection frequency in the possible relation task. Attested compound words were rated after self-paced reading presentation. (*right*) Main effect decomposed across all 16 random intercept effects for relation.

Interim discussion: We replicate the findings of analysis 1 for existing compounds, here rated in context and with a smaller sample size, demonstrating that they are robust to changes in the task, and that averaging across multiple contexts is not necessary to obtain good representations. Here, while semantic relations are implicit and thus unlikely to be attested in corpora, the compounds are not, and the contexts in which they occur in a model's training set might be sufficiently informative to drive the observed results. I argue that this consideration does not in itself undermine the present approach: Humans are also sensitive to the distributional regularities of word usage and meaning itself can be conceptualized, at least in part, in terms of distributional usage; this general notion naturally applies to the specific case of compound word processing and understanding (e.g., relational dominance effects, Gagné & Shoben, (1997); morphologically productive constituent families, Libben, (2014)). However, humans can often readily interpret novel combinations, and an indication of genuine combinatorial abilities would require models to generate good representations of both unseen combinations and their interpretations. Thus, relying on otherwise identical experimental conditions, I then tested whether the findings of analysis 2 generalize to the case of novel compound words.

5.3.3 Analysis 3: Novel compounds rated in context

Dependent variables: The dependent variable was the number of participants selecting a given interpretation for a given novel compound in Benjamin & Schmidtke (2023). Task procedure was identical to that of analysis 2 for existing compounds. 100 novel compounds were constructed by crossing all left and right constituents from the LADEC database (Gagné et al., 2019) and ensuring their absence in the Contemporary American English (COCA; Davies, 2008) and the Wikipedia corpus (Davies, 2015). 47 raters judged the 100 novel compounds embedded in 2 sentences each (21.5 median judgments per compound and sentence).

Model predictions: As in analysis 2, because novel compounds were presented in context, I generated CWEs for the original sentences presented to participants, computed relation probabilities for all layers and models, and standardized the results. The dataset comprised 2 sentences for 100 compounds, each with 36 paraphrases, resulting in 7400 sentences.

Statistical considerations: The statistical procedure was identical to that described in section 5.3.2.

Results: Once again, relation probability and selection frequency were positively related. Consistent with previous findings, BERT's fit increases almost monotonically, peaking at the 12th layer, while Llama-2-13b performance progresses in a U-shaped manner, peaking at layer 16 (figures A.3.5 and A.3.6). Compared to studies 1 and 2, the performance of both LLMs is significantly weaker. Indeed, the lowest AIC is achieved by the vector difference model, immediately followed by the additive model. BERT-base's top performing layer manages to outperform only the CAOSS model ($\Delta AIC = -5.4$), but not the vector difference and additive models ($\Delta AIC = 8.3$ and 4.8, respectively). All baseline models outperform Llama-2-13b layer 16 ($\Delta AIC = 31.8, 28.3, 18.1$). Moreover, fewer layers remain significant after applying Bonferroni correction, namely BERT-base layers 10, 11 and 12 and Llama-2-13b layers 14, 15, 16 and 17. Figure 5.5 shows how selection frequency varies as a function of relation probability for BERT-base layer 12 ($\beta = 0.22, p < .0001$). Results remain consistent after model criticism ($\beta = 0.26, p < .0001$) and are robust to alternative wording ($\beta = 0.19, p < .0001$). Weaker results for Llama-2-13b layer 16 were observed (i.e., $\beta = 0.20, p < .0001$), further weakened by model criticism ($\beta = 0.13, p < .0042$) but not alternative wording ($\beta = 0.21, p < .0001$).

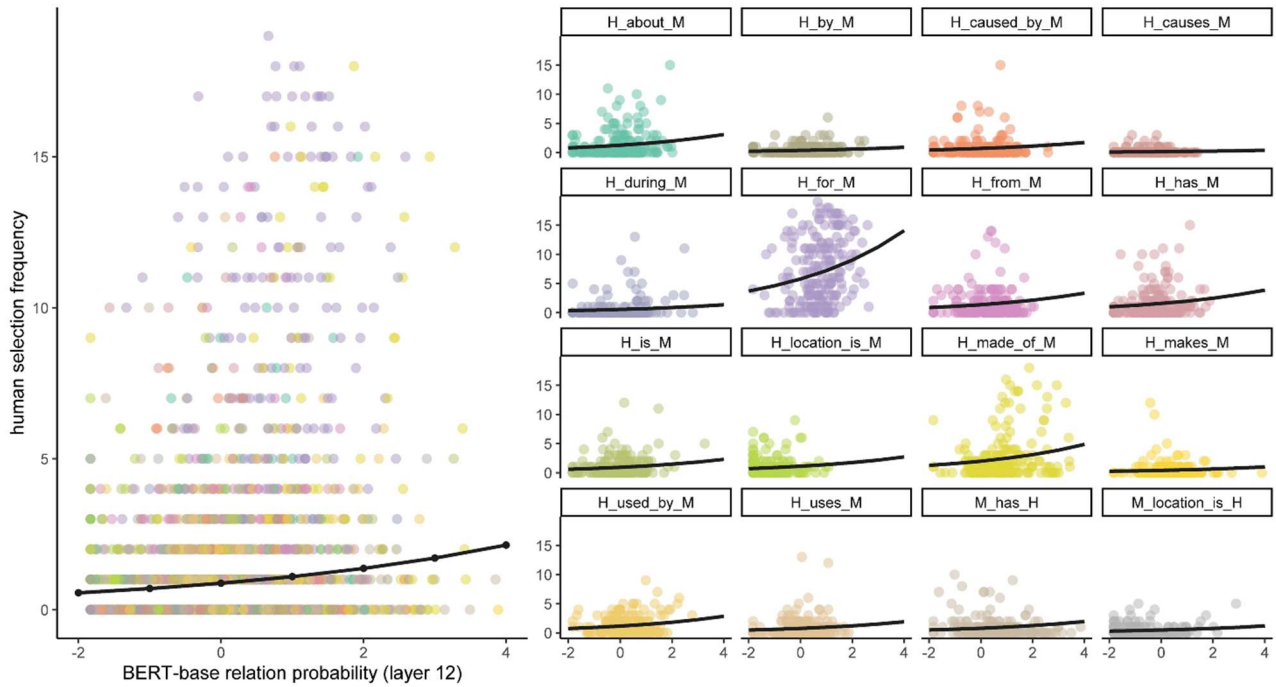

Figure 5.5: Relation probability effects for BERT-base layer 12. (*left*) Main effect of relation probability (scaled) on selection frequency in the possible relation task. Novel compound words were rated after self-paced reading presentation. (*right*) Main effect decomposed across all 16 random intercept effects for relation.

Interim discussion: In analysis 3, we find again a significant positive relation between relation probabilities, estimated in the space of some LLMs' layers, and selection frequency, a relation requiring adequate representations of both novel combinations *and* their implicit meanings. However, here LLMs' effects were considerably weaker, to the point that two baseline models outperformed even the best of BERT-base and Llama-2-13b layers. Because the data and analytical procedures employed are otherwise closely aligned to analysis 2, one can reasonably infer that the weaker effects mainly stem from the compounds' novelty. I will now discuss the findings more in detail.

5.4 Discussion

This chapter explored the ability of LLMs and cDSMs to reveal the implicit meaning of existing and novel compound words. Findings indicate that replacing a compound with a paraphrased version leads to changes in LLMs' contextualized word embeddings that are inversely proportional to the paraphrase's plausibility, as estimated by human raters. While this relation holds for both existing and novel compounds, results obtained for novel compounds are substantially weaker and below the baseline provided by simpler cDSMs.

Evaluation and comparison of semantic spaces. Consistent with previous findings (Buijtelaar & Pezzelle, 2023; Miletic & Schulte, 2023; Ormerod et al., 2024), LLMs show improvements over the cDSMs baselines for existing compounds rated in context and in isolation. However, turning to novel compounds, both BERT-base and Llama-2-13b are outperformed by all baseline models except CAOSS (figure A.5.6), suggesting that these LLMs struggle to meaningfully represent novel lexical combinations. Comparing model coefficients in analyses 2 and 3 (differing in the type of compounds rated – existing or novel – but otherwise identical in terms of sample size, task and analytical procedure), we see that cDSMs perform more consistently than LLMs. Specifically, comparing the beta coefficients estimated for existing and novel compounds, we observe a much more pronounced decrease in magnitude for LLMs compared to cDSMs. First, the consistent performance of cDSMs is in line with the successful application of (some of) these models to novel compound word semantics (Günther & Marelli, 2020, 2022). Second, the performance degradation of LLMs in the case of novel compounds aligns with recent NLP findings, which suggest that LLMs rely more strongly on constituent words (Yu & Ettinger, 2020, 2021) and familiar lexical items (Coil & Shwartz, 2023; Li et al., 2022). Here, BERT-base and Llama-2-13b might similarly suffer from an overreliance on training, an issue that would become especially apparent in the case of novel compound paraphrasing, since LLMs have none or extremely reduced exposure to the compounds *and* their paraphrases. It is also worth pointing out that, while CAOSS shows a slight but negligible advantage over the additive and vector difference cDSMs in analysis 1 ($\Delta AIC = 0.9$ and $\Delta AIC = 1.3$, respectively), it is substantially outperformed by both models in analysis 2 ($\Delta AIC = 30.4$ and $\Delta AIC = 5.5$; figure A.5.4) and in analysis 3 ($\Delta AIC = 10.2$ and $\Delta AIC = 13.7$; figure A.5.6). This goes against Günther & Marelli, (2022) demonstration that a linear probe can more accurately predict relational interpretations based on embeddings from CAOSS compared to the additive model, showing that the relational information it carries is encoded, at least in part, implicitly.

Because cDSMs are simpler class of models and are already capable of accounting for a wide range of compounding phenomena (Günther & Marelli, 2018; Marelli et al., 2017), improvements over a cDSM baseline are a necessary precondition to advocate for a shift to LLMs. Thus, these results do not support such a shift, at least for the specific dimension targeted. However, LLMs still offer a fundamentally novel modeling approach to context-dependent and token-level representations (Cassani et al., 2023; Pavlick, 2022). If these characteristics are prized, the positive findings across all conditions align with other evaluations (Buijtelaar & Pezzelle, 2023; Miletic & Schulte, 2023; Ormerod et al., 2024) in demonstrating LLMs to be at least viable representational models of compound word meaning.

The role of model depth: The representational adequacy of BERT-base grows almost always monotonically with model depth, peaking at the 12th and last layer (figures A.3.2, A.3.4, and A.3.6). This profile is consistent with Buijtelaar & Pezzelle, (2023) study of semantic transparency (but see Miletic & Schulte, 2023) and headedness. Indeed, since the successful interpretation of a compound depends on the correct assignment of head and modifier roles to its constituents, I speculate that the layer-wise increase we observe might reflect an underlying increase in the representation of headedness. Another tentative explanation rests on the role of context, namely the fact that contextualization magnifies the difference between plausible and implausible replacements in the extent to which they fit in the portion of semantic space occupied by the expression they replace (Cassani et al., 2023). When looking at Llama-2-13b, instead, a U-shaped pattern emerges (figures A.3.2, A.3.4, and A.3.6), suggesting that, for this model, contextualization is not monotonically beneficial. However, one should be careful when interpreting Llama-2-13b since much less research has been conducted on its representations, especially compared to BERT.

(Implicit) semantic relations in LLMs: While LLMs appear successful in modeling relational information (Piantadosi & Hill, 2022), one would like to know not only *how much* but also *how* two concepts are related. Indeed, Lake & Murphy, (2021) criticize LLMs for relying on general similarity metrics that conflate specific sources of relational information such as hyperonymy, parthood, location, and so on. The present approach still relies on cosine similarity, but shifts the perspective from comparing words to word *senses*. By keeping the context constant, changes in cosine similarity can be more directly attributed to target semantic relations. First, this makes this approach flexible, because it does not assume that the set of chosen semantic relations is fixed (see alternative wordings; table 5.1). Moreover, compared to other approaches based on probing (Günther & Marelli, 2022), no additional training is required to interpret results. Thus, at least in the domain of compounding, I claim that this approach addresses the criticism raised by Lake & Murphy, (2021) by offering an LLMs-based operationalization of *specific* relational information that is at once continuous, flexible, data-driven and nearly unsupervised (i.e., not involving any training but merely requiring the choice of target relations of interest). Overall, by focusing on the internal structure of compounds, I contend that these findings support both the notion that LLMs represent compound word semantics and relational information beyond general similarity.

General considerations and future developments: Overall, the present approach is consistent with recent theoretical proposals in cognitive psychology and psycholinguistics which posit that word meaning should not be conceptualized as a particle-like entry in a mental lexicon, but rather as a set of possibilities that might or might not be realized in a given instance of language use (Libben, 2017;

Rodd, 2020). This notion is consistent with the CARIN and RICE theories, which propose that compound word understanding involves the evaluation of multiple senses that are at first only possible, and that compete to be selected and realized. In this context, LLMs provide representations of those meanings that are both possible and implicit. The compound CWE corresponds to the meaning of a combination that has not yet been interpreted, corresponding to an initial ambiguous or "blend state", whereas the CWE of a paraphrase describes the "stable state" where a specific interpretation obtains (Rodd, 2020). Moving away from the blend state, closer CWE are encountered first, corresponding to paraphrases with higher selection probability. Because this study adopts a strictly representational approach (i.e., it focuses on the degree of alignment between the structure of artificial and human conceptual systems) future studies should attempt to explain how representations are developed, and how they influence behavior in time. Relatedly, all analyses were limited to high-level, deliberate judgments, while previous studies show that the conflict among possible relational interpretations impacts word processing (Benjamin & Schmidtke, 2023; Schmidtke et al., 2016; Schmidtke, Gagné, Kuperman, & Tucker, 2018). Thus, future studies could target processing dimensions, first by testing relation probability against chronometric data, and second (and more interestingly) by leveraging LLMs estimates that are compound-specific and driven by data (e.g., BERT's CWE capture words' clusterability into senses; Soler & Apidianaki, (2021)), sidestepping the problem of relying on a fixed taxonomy of semantic relations (Downing, 1977; Gagné & Shoben, 1997).

Chapter 6

General discussion

6.1 Summary of contributions

The aim of this dissertation was to demonstrate that distributional models of semantics, by formalizing both the representation of individual concepts and the functions that combine them, are a valuable tool to investigate the implicit information that characterizes conceptual combination and that distinguishes it from "overt" forms of combination, such as sentences. On the one hand, by leveraging computational models, this work put to the test various claims about how conceptual combinations are represented and processed. On the other hand, the latter provided a testing ground for computational models themselves. More specifically, this work focused on modeling different routes to conceptual combination characterized by (i) different representations of the same constituent concept (morphological transcendence; Libben, 2014), (ii) different representations of the same complex concept (lexicalized, combinatorial; Gagné & Shoben, 1997; Libben, 2010), and (iii) different sources of information on which these representations are based (language or multimodal information; Lynott & Connell, 2010). Although varying in scope and domain, all point to the notion that combination rests on a set of implicit representations going beyond the surface-level modifier, head, and complex concept. This raises the possibility that successful combination is accomplished by leveraging all these multiple routes, following a strategy of "maximization of opportunity" (Libben, 2010): constituent concepts are updated based on their morphological history (Libben, 2014), processes of retrieval and active combination are automatically enacted (Günther & Marelli, 2020; Libben, 2010), and determined by both language and multimodal simulation (Lynott & Connell, 2010), possibly resulting in multiple competing interpretations (Gagné & Shoben, 1997; Schmidtke et al., 2016; Spalding et al., 2010). I now provide a summary of the main findings across the four studies that comprise this dissertation and discuss how they relate to the above matters.

(i) Chapter 2 supports the notion that the same concept can be represented differently in relation to its role as a (modifier) constituent. Results show that more experienced readers become progressively less reliant on the original, free-word meaning of modifier constituents, while the benefit of semantic transparency (or "semantic boost"; Schmidtke et al., 2018) increases for compounds of higher frequency, consistent with the idea that compound frequency reflects

combinatorial knowledge (Baayen et al., 2010). Moreover, results confirm the superiority of compositionally-derived measures of transparency (Amenta et al., 2020; Günther & Marelli, 2018), supporting the notion that processing compound words entails more than the free-word, lexicalized meaning of constituents and compounds. These findings are consistent with the morphological transcendence (Libben, 2014) and lexical quality (Perfetti, 2007; Perfetti & Hart, 2002) hypotheses, which similarly posits the progressive specification of a word's meaning in relation to its constituent-specific role. They are also compatible with an account of automatic conceptual combination, whereby the application of CAOSS's *M* and *H* matrices corresponds to the active generation of as-constituent representations guided by previous experience with compounding (Günther & Marelli, 2020). More broadly, results are consistent with evidence showing that the semantic characteristics of compound words are influenced by reading experience (Schmidtke, Gagné, Kuperman, & Spalding, 2018), which modulates their effects on behavior (Pham & Baayen, 2013; Schmidtke, Van Dyke, et al., 2018). Finally, it is worth noting that both interactions were observed only for modifiers; analogous measures for head constituents did not significantly improve model fit despite targeting gaze duration, where information about heads is already available. Such constituent asymmetry is consistent with previous computational studies of semantic transparency using the CAOSS model (Günther, Petilli, & Marelli, 2020; Günther & Marelli, 2018, 2020) and with the CARIN and RICE claim that combinatorial (relational) information is provided by the modifier rather than the head constituent (Gagné & Shoben, 1997; Jones et al., 2008; Spalding et al., 2010).

(ii) Chapters 2, 4, and, in part, 5 dealt with whether multiple types of the same combination might be represented and processed simultaneously via the retrieval of lexicalized meaning or by active combination.

In Chapter 2, measures of semantic transparency and consistency that relied either on the lexicalized meaning of familiar compounds or their compositional representation (as generated by CAOSS), were compared. Findings replicate the superiority of compositionally-derived measures of transparency, although with a difference: rather than linking free-word constituents to the compositional compound representation (the approach followed in previous studies of this type; Günther, Marelli, et al., 2020; Günther & Marelli, 2018, 2020), the best-fitting predictor was defined based on the similarity between as-constituents and the *lexicalized* compound (figure 2.1). I speculate that this difference is due to task requirements. Previous studies focused on lexical decision tasks and timed sensibility judgments, tasks that can all be carried out without accessing lexicalized information. Instead, participants in CompLex (i.e., the resource considered in Chapter 2; Schmidtke et al., 2021) were asked to read sentences for comprehension, thus arguably relying on the actual, lexicalized compound meaning rather than its compositional approximation. Günther, Marelli, et al.,

(2020) advanced analogous considerations in a study of German compounds, where compositional measures predicted lexical decision times but not eye movements, which were instead predicted by non-compositional measures linking the free-word meaning of both constituents and compounds. Integrating the findings of Chapter 2 with previous studies, the emerging picture suggests that automatic combination is always enacted but is at the service of the task at hand, which might prize more the recollection of lexicalized compound meaning or the formation of a transient compositional one.

In Chapter 4, different cDSMs were used as theoretical models of the semantic and compositional representations supported by brain regions of interest in four published fMRI studies, tested in confirmatory analyses of *a-priori* ROIs (i.e., LATL, LAG, RAG, and LIFG; Coutanche et al., 2018), and in exploratory analyses conducted within a mask of the general semantic network (Jackson, 2021). The main result of Chapter 4 is that confirmatory and exploratory analyses converge on a cluster of representations of the multiplicative cDSM in *pars triangularis* of the LIFG (BA45), running against the claims that this region is only univariately modulated by task complexity, and qualifying the type of combination as (feature) conjunction or intersection (Baron & Osherson, 2011; Baroni, 2013). Importantly, multiplicative representations were still observed in a confirmatory analysis targeting LIFG of Graves et al., (2010; experiment 1) data, wherein participants performed a form-based 1–back task that did not require access to any semantic (and thus, compositional) information. As with Chapter 2, this finding is consistent with a combinatorial system that is automatically triggered by relevant stimuli and expands previous neuroimaging findings of automatic semantic activation to the case of composition (Liuzzi et al., 2021). However, unlike the behavioral studies discussed so far, the CAOSS model never approached significance in any of the analyses performed in Chapter 4, suggesting that the representations provided by this model, while informative for explaining behavior, fall short of modeling neural representations (at least with RSA; Kriegeskorte et al., 2008). However, care should be taken in coming to this conclusion, since the re-analyzed data came from four fMRI studies differing in task and stimuli, and thus, a possible explanation for the lack of findings is that only a more basic form of combination was shared across these studies and thus detectable in their aggregate. This also conforms to the considerations advanced above, namely that the relevance of different representations provided by the CAOSS framework (figure 1.3.2) depends on task requirements. Finally, Chapter 4 directly tackled the issue of simultaneous retrieval and combination by focusing on a subset of 45 subjects reading 45 familiar compound words, which were represented both as lexicalized units (defined by their free-word word2vec embeddings) or compositionally (the outcome of cDSMs). Assuming that participants would engage in both retrieval and combination (Libben, 2010), I expected to observe clusters of both lexicalized and combinatorial

representations, possibly localized in different regions of the general semantic network. However, no significant clusters of representational similarities were observed, neither of the lexicalized nor of the compositional models. The latter contradicts the evidence for clusters of multiplicative representations in BA45 and left middle-superior temporal lobe, which were observed in the previous exploratory analysis that included all participants and stimuli but was otherwise identical. The most likely explanation of this inconsistency is that the searchlight analysis focusing on familiar compounds simply lacked the statistical Power to detect representational similarities based on semantic models. Therefore, evidence for simultaneous retrieval and combination remains inconclusive.

Chapter 5 showed that information about implicit semantic relations can be recovered by the compound-paraphrase similarities in the semantic spaces of two LLMs (BERT, Devlin et al., 2019; Llama-2-13b, Touvron et al., 2023) and, albeit to a lesser extent in the case of existing compounds, of (c)DSMs (CAOSS, Marelli et al., 2017; vector difference, Mikolov, 2013; Levy & Goldberg, 2014; vector addition, Mitchell & Lapata, 2010). Thus, Chapter 5 shows that distributional information is at least sufficient to *represent* the plausibility of multiple (relational) interpretations of the same complex concept. This is consistent with CARIN and RICE's proposal that conceptual combination entails the representation of multiple possible semantic relations (Gagné & Shoben, 1997). Once represented in a semantic space, information about these candidate relations might influence processing as theorized by Gagné & Shoben, (1997) and shown empirically with human ratings (Benjamin & Schmidtke, 2023; Schmidtke et al., 2016; Schmidtke, Gagné, Kuperman, & Tucker, 2018).

(iii) In Chapter 3, participants selected which of two images best depicted their visual interpretation of novel compound words. Here, image information was represented in the vision spaces of convolutional neural networks (CNNs; Chatfield et al., 2014; Günther et al., 2022, 2023), while participants' visual interpretation was formalized by "predicted prototypes", namely the result of the sequential application of combinatorial operations conducted in language space (implemented by the CAOSS model; Marelli et al., 2017), and of language-to-vision mapping (Günther, Petilli, Vergallito, et al., 2020; Sorscher et al., 2022). Predicted prototypes captured untimed choices of compound interpretation (i.e., destructive vs. non-destructive interpretation), but not those of image typicality (i.e., which image best depicted one interpretation). Moreover, mapped prototypes predicted image selection times, albeit only in accurate compound-head trials. Notably, the language-to-vision mapping based on the standalone head concept consistently led to the best model fit, outperforming CAOSS and representing the only model whose predictions significantly aligned with

judgements of image typicality. First, these results support the idea that the second-order similarity or "systems alignment" (Aho et al., 2022) between distributional and vision models can be leveraged by humans to bootstrap information about referents from how these are talked about in natural language (Aho et al., 2023; Louwerse, 2018), exploiting the representations available in one system (the distributional representation of constituents) to infer information lacking in another (the perceptual features of their combination). Second, the superior performance of the head constituent and the lack of effects on reaction times in some conditions suggest that, while language likely played a part in participants' behavior, additional combinatorial processes were at play, possibly outside the language domain. This consideration conforms to the ECCo model (Lynott & Connell, 2010), positing a fast, "shallower" language route (here, provided by the distributional meaning of the head), complemented by "deeper" combination operating directly on sensorimotor information (i.e., "situated simulation"; Lynott & Connell, 2010; Wu & Barsalou, 2009). A stricter test of this conclusion would require the direct comparison of language-based and vision-based compositional representations (e.g., Günther, Petilli, & Marelli, 2020), together with the exploration of different stimulus onset asynchronies in the experimental paradigm.

6.2 How can we model conceptual combination? Comparing different solutions

Different computational models of conceptual combination were considered throughout this dissertation, namely cDSMs based on arithmetic operations (i.e., additive and multiplicative models; Chapter 4 and 5; Mitchell & Lapata, 2010), cDSMs based on "compositional" function application (i.e., the CAOSS model; Chapter 2, 3, 4 and 5; Marelli et al., 2017), and LLMs (i.e., BERT-base and Llama-2-13b; Chapter 5; Devlin et al., 2019; Touvron et al., 2023). The phenomena considered here are too limited to draw general conclusions about the *performance* of different model classes. Indeed, LLMs were considered only in Chapter 5, and were compared to cDSMs only in relation to their ability to represent implicit relational information. Nonetheless, besides performance, the studies presented help shedding light on the theoretical merits and drawbacks of each approach.

Generally speaking, cDSMs and LLMs approach semantic combination from opposite directions (Pavlick, 2022): cDSMs take a bottom-up approach whereby compositional functions operate on simpler arguments to build up complex meanings, while LLMs seek a top-down explanation that characterizes the meaning of individual tokens in light of the role they play within

the broader sentential context. This difference has profound implications for modeling conceptual combination. First of all, a bottom-up approach makes model comparison more interpretable and controllable. Indeed, as shown in Chapters 2 and 4, cDSMs offer alternative operationalizations of conceptual combination via different compositional functions. When cDSMs differ exclusively along the combinatorial dimension, the outcome of model comparisons can be transparently attributed to the compositional functions which, if interpretable, can be related to (neuro)cognitive theories of conceptual combination; this was the rationale behind the tests of ROI-model correspondence performed in Chapter 4. Instead, in LLMs there is no qualitative difference between simple and complex words, the only (quantitative) difference being that the former are more likely to be decomposed into more token components, which the model learns to link via attentional mechanisms. In this sense, complex words can be considered as a step in a continuum of complexity that starts from single tokens and extends to longer multitoken expressions and even sentences. This makes LLMs' approach appealing for at least two reasons. First, it is parsimonious, as it relies on general combinatorial operations based on attention (Vaswani et al., 2017) and does not require additional machinery to carry out special cases such as compounding (e.g., CAOSS's *M* and *H* matrices; Marelli et al., 2017). Relatedly, it allows for greater representational flexibility by relying on a single framework for representing linguistic units of different levels of complexity, whereas cDSMs run into various limitations when attempting to scale up composition to longer sequences of text (see general introduction, section 1.3.3). However, the way in which constituents are combined is less interpretable and controllable, whereas explaining conceptual combination should be, in its core, a quest for characterizing combinatorial transformations (Chapter 4). Taking Chapter 4 as an example to illustrate the issue, it is difficult to see how the same study could have been conducted using LLMs, given the need for alternative, interpretable combinatorial representations for the same complex words. While overt semantic interpretations can be used to characterize alternative outcomes of conceptual combination (as in Chapter 5), ultimately one would want to know how these are covertly encoded by a compound, that is, how the representations of two concepts are combined in a way that approximates an overt description. This problem is perhaps even more pronounced when comparing lexicalized and compositional meanings of the same complex word (Chapter 2 and 3). While straightforward for (c)DSMs, LLMs cannot distinguish between the two since they do not actually "store" meanings, either prototypical or exemplar; rather, they generate embeddings that tend to preserve a core of "prototypical" or "decontextualized" meaning modulated by context (Cassani et al., 2023). Different contexts might be biased toward lexicalized or compositional interpretations. For example, consider the opaque compound "buttercup", which speakers learn is a type of flower, thus building a corresponding lexicalized representation. By designing appropriate contexts, it might be

possible to induce a compositional, literal reading of "buttercup", e.g., in: "The leftover **buttercups** were greasy" compared to a control such as: "The delicate **buttercups** are yellow". While interesting, this approach only postpones the upstream issue, as it is still possible to conceive lexicalized readings of the former and compositional readings of the latter, and LLMs should thus still need to represent alternative meanings for the same compound in the same context.

On the one hand, one could see in the above a testament to the limitations of the top-down approach followed by LLMs. However, the bottom-up approach to conceptual combination characterized by the classic principle of compositionality is not the only account available. For example, Ramscar, (2019) argues for rethinking language in terms of information theory (Shannon, 1949), i.e., an exchange of information whereby the goal of the sender is reducing the uncertainty of the receiver about the intended meaning encoded by a message. On this account, language is not about the *inductive* combination of independent parts, but a *deductive* process of meaning discrimination (i.e., elimination of alternatives). In this context, by re-using preexisting words, compounding can be considered a form of "encoding parsimony", consistent with Zipf's "principle of least effort" (Zipf, 1949) and Piantadosi et al., (2012) account of ambiguity as the result of a trade-off between the communicative pressures of *clarity* and *ease.* As for ambiguity, the benefit of compounding lies in the possibility of sending "easier" messages by re-using short and frequent words. For ambiguity, these benefits come at the cost of clarity, in the occasional errors of disambiguation (which, Piantadosi et al. argue, are mostly avoided by contextualization). The cost of compounding, instead, likely comes from the need to infer implicit information, which is also often ambiguous (Schmidtke et al., 2016) and dependent on context (Benjamin & Schmidtke, 2023). In this scenario, LLMs could be considered models of such disambiguation, and the results of Chapter 5 could be seen from this perspective: the fact that the distance between a compound and its paraphrase is proportional to the paraphrase's plausibility suggests that, if one were to reconstruct the intended meaning of compound from neighboring interpretations, more plausible messages would be recovered with more ease. This consideration is in principle consistent with connectionists models of ambiguity resolution (Rodd, 2020), whereby disambiguation consists in moving away from a "blended" state into one of the familiar interpretations of the word, and with CARIN and RICE proposals of the competition among relational interpretations (Gagné & Shoben, 1997; Schmidtke et al., 2016; Spalding et al., 2010).

Overall, setting aside the issue of cognitive plausibility (see general introduction), the choice of cDSMs or LLMs rests on both theoretical and methodological considerations. Theoretically, cDSMs follow a tradition where conceptual combination is the outcome of bottom-up (Pavlick, 2022), "inductive" (Ramscar, 2019) function application (Baroni et al., 2014; Baroni & Zamparelli, 2010),

whereas LLMs take a top-down perspective (Pavlick, 2022) amenable to be subsumed into a general "deductive" process of meaning disambiguation (Ramscar, 2019; Rodd, 2020) and of selection of competing interpretations (Gagné & Shoben, 1997; Spalding et al., 2010). Besides theoretical commitments, cDSMs and LLMs come with different methodological strengths and weaknesses. cDSMs allow for the controlled generation of alternative representations of the same complex concept (constituents and as-constituents, lexicalized and compositional), and a rich system of representations supporting nuanced interpretations (e.g., different operationalizations of semantic transparency; Chapter 2). Instead, the representation of a complex concept is intertwined with its context in LLMs, which take an "exemplar" approach (while generating prototypes is still possible by averaging across contexts; Bommasani et al., 2020; Cassani et al., 2023; Chapter 5) and can flexibly handle longer sequences of text, placing overt and covert combinations on a similar standing.

6.3 Future work

There are several new developments that follow from the studies included in this dissertation and their limitations.

The morphological transcendence hypothesis (Libben, 2014) and the lexical quality hypothesis (Perfetti & Hart, 2002) are essentially based on learning new lexical representations or refining existing ones. However, CAOSS is quite limited in modeling how this learning unfolds. Indeed, one first assumes the availability of constituent representations that are the outcome of a plausible learning model (i.e., word2vec; Mandera et al., 2017). Only then is CAOSS grafted on top of these representations, and its parameters (the M and H matrices) trained on them. Thus, this approach separates the learning of conceptual combination to the learning of concepts themselves, while the two should plausibly go hand in hand. Future studies should focus on bringing together constituents and combinations within the same learning system.

A natural extension of the framework presented in Chapter 3 is the inclusion of a model of visual composition, e.g., a CAOSS architecture that directly operates in vision space by combining CNNs embeddings (Günther, Petilli, & Marelli, 2020). Because Chapter 3 explored the extent to which responses to visual combinations could be accounted for by language combination and the systematic mapping that links language and vision, stimuli were chosen with respect to language predictions (i.e., language predictions defined which images to employ as the most and least prototypical), and so a direct comparison between vision-based and language-based combination

models would be biased in favor of the former. Thus, future studies should focus on an unbiased comparison between the two models.

Turning to vision models, CNNs were chosen because of the available cognitive evaluations (Günther et al., 2022) and model comparisons (Günther et al., 2023), but first and foremost because they received relatively little language supervision, so that the performance of vision or language combination models can be more safely attributed to the respective modality rather than the shared structure brought by stronger supervision. Nonetheless, motivated by the fact that words are also learned and experienced in contingency with their referents, recent studies have shown that multimodal models pre-trained to align language and vision embeddings in the same multimodal space (e.g., CLIP; Radford et al., 2021) offer promising models of word learning (Vong et al., 2024), object properties (Walbrin et al., 2024), and of the brain alignment of language and visual systems (Tang et al., 2023) over and above their unimodal counterparts (Wang et al., 2023). Thus, future studies could consider investigating the interplay between language and vision combination using multimodal representations (e.g., extending Chapter 3 and 4 by exploring the alignment of possible relational interpretations and synthetic images).

Turning to the issue of the cognitive neuroscience of conceptual combination, Chapter 4 suffers from methodological drawbacks related to the fact that the data were not matched to the research questions and methodologies followed in their re-analysis (e.g., signal dropout in LATL). Thus, an obvious development in this direction is to optimize data collection, from signal sampling to stimuli definition. For example, concerning the stimuli, one would seek to employ only familiar compounds, so that lexicalized and compositional representations would be available for all stimuli. Finally, conceptual combination studies have relied almost exclusively on linguistic stimuli (but see Baron & Osherson, 2011; Coutanche & Thompson-schill, (2015) for simple visual stimuli). However, because conceptual combination likely relies on sensorimotor information (Günther, Petilli, & Marelli, 2020; Lynott & Connell, 2010), this raises the question of whether the same network of ROIs is also involved in *visual* conceptual combination. Evidence suggests that, at least for LAG and LATL, this should be the case: these ROIs have been long identified as “multimodal” or “amodal” regions (Damasio’s convergence zones; the hub-and-spoke model of semantic memory; Ralph et al., 2016), sitting atop of a sensorimotor-abstract representational hierarchy (Margulies et al., 2016). Moreover, the apparent division of labor between attributive and relational combinations in LATL (conceptual specificity) and LAG, respectively (Boylan et al., 2017), naturally translates into the analogous visual categories of destructive and non-destructive interpretations (Lynott & Connell, 2010). Therefore,

future studies could focus on complex visual combinations, with the images developed in Chapter 3 offering suitable stimuli.

Finally, Chapter 5 adopted a representational approach, focusing on the degree of alignment between model representations and human judgements. Thus, on the modeling side, future efforts could be devoted to explore how representations are developed and how they influence behavior, e.g., drawing on the parallels between transformers and deep associative networks (Ramsauer et al., 2021). On the empirical side, the natural development of this research line is to shift the focus from high-level, deliberate judgments, such as those of the possible relation task, to more automatic forms of semantic processing. Indeed, possible relation task judgments have been used to estimate the degree of competition among relational interpretations theorized by CARIN and RICE, estimates which were shown to impact word processing (Benjamin & Schmidtke, 2023; Schmidtke et al., 2016; Schmidtke, Gagné, Kuperman, & Tucker, 2018). Here, model-based relation probability measures could be tested against chronometric data to replicate findings obtained using participants' judgments. Second, and more interestingly, LLMs representations could be leveraged to generate data-driven predictions that would otherwise be prohibitive to obtain. For example, LLMs could sidestep the problem of coming up with an *a priori* comprehensive taxonomy of possible interpretations (Downing, 1977; Gagné & Shoben, 1997) by generating compound-specific estimates based on how each compound is actually used in corpora (e.g., Soler & Apidianaki, (2021) show that BERT's embeddings capture words' clusterability into senses).

Clearly, a lot of work still needs to be done to understand how the multiple routes of morphological transcendence, multimodality, retrieval and active composition each contribute (if they indeed do) to conceptual combination. In this dissertation, not all dimensions emerged as clearly as initially hypothesized – if they were evident at all. Indeed, evidence for the simultaneous engagement of retrieval and (possibly, multiple) combination processes remains inconclusive. Instead, other findings, such as the influence of language on visual combinations, represent what I believe to be initial but promising steps toward a multimodal theory of conceptual combination. Yet other results (e.g., the impact of morphologically transcended units, the automaticity of composition) corroborate, qualify and extend previous work. Finally, it is also important to note that, as task demands likely play a significant role in determining the degree of engagement of each one of these routes, one should not expect all of them to be always at play. This observation raises what is arguably a bigger challenge, namely how these routes should interact among each other as part of the general faculty of conceptual combination. What the present dissertation suggests is that, for many of these factors to be meaningfully defined and testable, they should be formalized, and in order to better inform one

another, they should be formalized within a single model (e.g., CAOSS) or modeling framework (e.g., cDSMs), where they eventually might lead to a systematic and cohesive theory arising from their mutual constraints.

Appendix

Appendix A.1

A.1.1 Validation concepts: selection and synthetic images generation

Validation concepts selection: Of the 1854 concepts in THINGSplus, 1101 are available in ImageNet-21k but not in ImageNet-1k. From these, a subset of compound and simple concepts to be used in the validation phase was selected. First, all available 217 compound words were considered, keeping those for which at least 13 images of size greater than CNNs' model input size were available in ImageNet-21k. Hence, five compound words were discarded ("doorstop", N = 8; "dumbwaiter", N = 1; "handbrake", N = 1; "pitchfork", N = 3; "springboard", N = 8), resulting in 212 compound words. The raw frequency of their constituents in the Baroni et al., (2014) corpus did not exceed 50. Then, an additional set of 212 single words was considered. The same criteria were applied, privileging words that were used as constituents in any attested compound. Therefore, the final validation set comprised 424 concepts (212 compound words and 212 simple words); all had a concreteness rating above 4 in Brysbaert et al., (2014) and were represented by at least 13 images whose size exceeded the largest CNN input size (i.e., $256 \times 256 \times 3$, DarkNet-53) in both THINGSplus and ImageNet-21k.

Generation of synthetic images: The objective was to obtain 3 realistic images for each concept. OpenAI DALLE-3 (Betker et al., 2023), a large diffusion model (Zhang et al., 2023), was used to generate sythetic images. DALLE-3 was chosen simply as a tool to generate realistic synthetic images based on complex prompts, without applying any cognitive modelling considerations. In a preliminary investigation, it was found that, for most concepts, at least 1/3 of generated images were realistic. Thus, 9 synthetic images of each concept were generated (style = "natural", image size = 1024×1024) based on the following prompts:

1) Two prompts stating: "A photo of a/an [CONCEPT NAME].";
2) Two prompts stating: "A photo of a/an [CONCEPT NAME], [LOCATION].", where [LOCATION] was defined manually by the first author.
3) Two prompts stating: "A photo of a/an [CONCEPT NAME], [DESCRIPTION].", where [DESCRIPTION] was the output of GPT-3.5-turbo-instruct (max tokens = 80, temperature = 0.7) prompted with the following: "Describe the visual appearance of [CONCEPT NAME], using less than 15 words. Begin the sentence with: [CONCEPT NAME] is ...";

4) Two prompts stating: “A photo of a/an [CONCEPT NAME], [DESCRIPTION].”, where [DESCRIPTION] was the output of GPT-3.5-turbo-instruct (max tokens = 80, temperature = 0.7) prompted with the following: “Describe how [CONCEPT NAME] looks like, using less than 15 words. Begin the sentence with: [CONCEPT NAME] is ...”;
5) One prompt stating: “A photo of a/an [CONCEPT NAME], [DESCRIPTION].”, where [DESCRIPTION] refers to the WordNet or Wikipedia description of the concept provided by THINGSplus.

Because DALLE-3 normally rewrites its prompt, prompts were also prefixed with the following instruction: “I NEED to test how the tool works with extremely simple prompts. DO NOT add any detail, just use it AS-IS:”. Some images were not generated on the first attempt due to content policy violations. In these cases, prompts were rewritten manually until an image was successfully generated. At most two rewriting attempts proved to be enough to obtain all images, with the exception of the “pantyhose” and “underwear” concepts, which consistently violated DALLE-3 content policy and were removed from the validation set. Finally, following their WordNet definition, prompts for 6 ambiguous words (“bank”, “cap”, “honeypot”, “horn”, “mouthpiece”, and “stamp”) were manually rewritten. After image generation, 3 out of every 9 synthetic images were selected. For an image to be considered valid, it had to depict at least partially, but always recognizably, the target concept in a realistic (i.e., photographic) style. When more than 3 images were deemed valid, images with complex, realistic backgrounds were preferred, as well as those that differed the most among each other, in order to increase their heterogeneity. Of the 422 concepts, 404 were judged to have at least 3 valid images (407 simple words and 197 compound words) and constituted the final validation set.

A.1.2 Parameter selection of the elastic net regularization

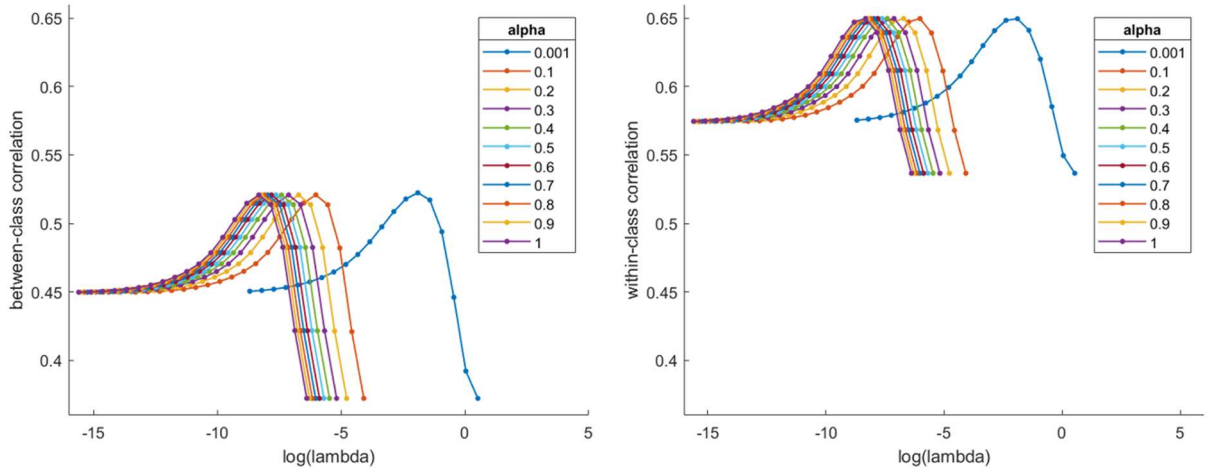

Figure A.1.1: Parameter selection of the linear regression model with elastic net regularization. Performance of the linear regression model on the validation set as a function of combinations of α and λ parameters.

A.1.3 Scores of the Vividness of Visual Imagery Questionnaire (VVIQ)

The reason for administering the VVIQ (Marks, 1973) was to identify and exclude participants falling below the threshold for aphantasia (VVIQ < 24; Zeman et al., 2020). For 5 participants, the response to some questions was not available; it these cases, the lowest score for these items was assumed.

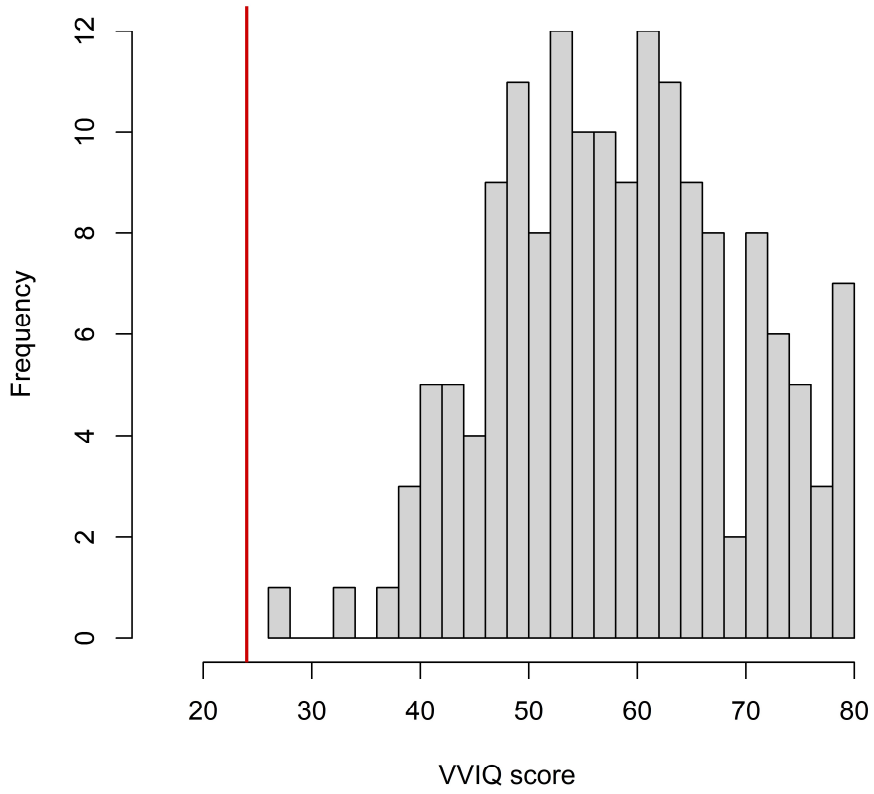

Figure A.1.2: Frequency distribution of VVIQ scores (N = 160).

A.1.4 Robustness analysis across reaction times thresholds

In the preregistration of these analyses, I set out to use an upper threshold of 4000 msec or 3.5 SD above the mean, whichever proved lower. Because the choice of the upper reaction time threshold was partly arbitrary, I checked whether results are robust to different RTs limits. Thus, statistical models were re-fitted using different thresholds at constant log-transformed RTs intervals starting at 3000 msec and passing through 4000 msec (Figure A.1.2). Results proved to be robust across different RT thresholds (Table A.1.1).

Table A.1.1: Analysis 3 model results across different RT thresholds.

Upper RT threshold	datapoints	coefficient type	β	p
3	4767	Intercept	0.24	<.0001
		diff. similarity	-0.23	.009
		interpretation	0.06	<.0001
2.46	4919	Intercept	0.27	<.0001
		diff. similarity	-0.21	.019
		interpretation	0.06	<.0001
4	5029	Intercept	0.29	<.0001
		diff. similarity	-0.24	.011
		interpretation	0.06	<.0001
4.62	5100	Intercept	0.30	<.0001
		diff. similarity	-0.24	.012
		interpretation	0.06	<.0001
5.33	5151	Intercept	0.31	<.0001
		diff. similarity	-0.25	.010
		interpretation	0.06	<.0001
6.16	5180	Intercept	0.31	<.0001
		diff. similarity	-0.27	.006
		interpretation	0.06	<.0001
7.11	5204	Intercept	0.32	<.0001
		diff. similarity	-0.27	.006
		interpretation	0.06	<.0001

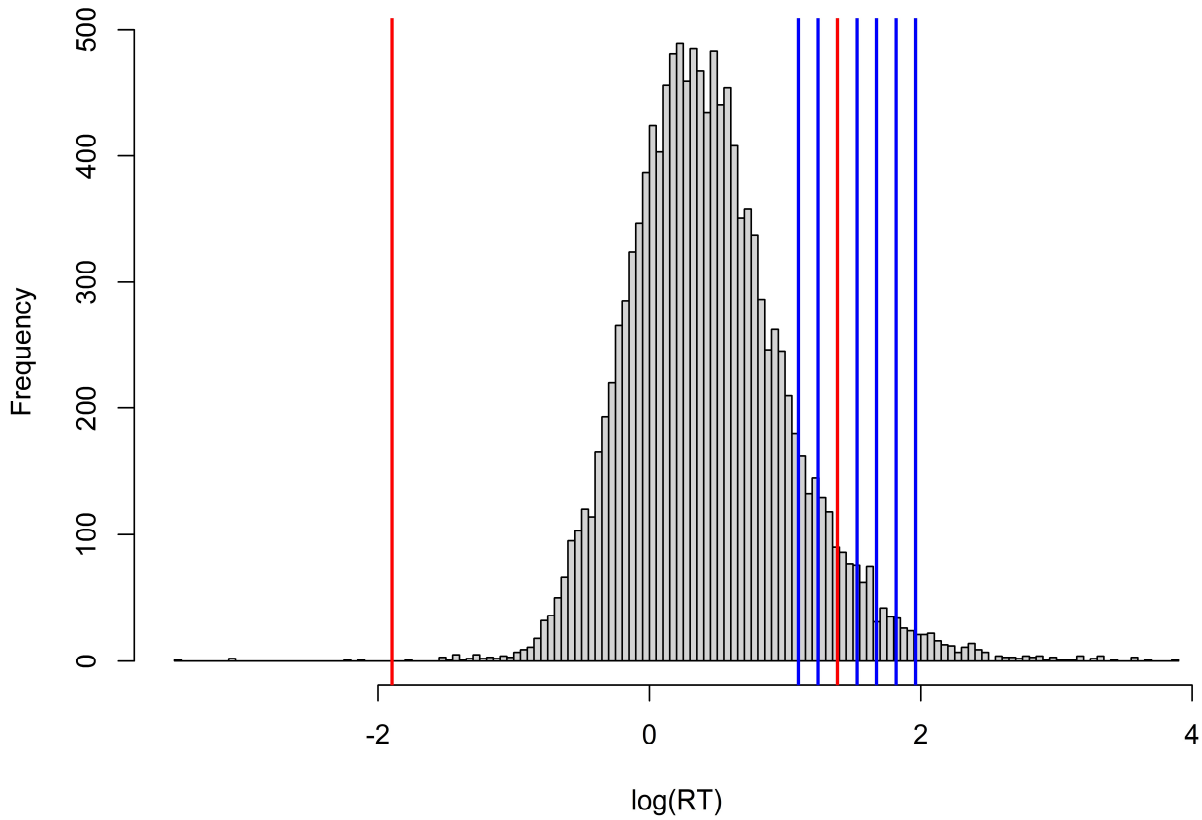

Figure A.1.3: Frequency distribution of log-transformed RTs from the whole dataset. Red lines indicate the lower (150 msec) and upper (4000 msec) RT limits used in analysis 3. Blue lines indicate the new upper threshold used.

A.1.5 Model comparison

The following tables show additional information about model comparison for each analysis. ΔAIC is the difference in the Akaike information criterion between a model and the best-fitting model. Akaike weights indicate the strength of evidence for a given model with respect to all models considered. The rightmost columns report the ratio between the Akaike weights of a target model and the best-fitting model (i.e., the head), and the probability that the latter is to be preferred over the former (Wagenmakers & Farrell, 2004; e.g., in analysis 1, the head concept is 19.4 times more likely than CAOSS to be the best model).

Table A.1.2: Results of language model comparison in analysis 1.

model	ΔAIC	**Akaike Weight**	**Weight ratio**	**Normalized probability**

CAOSS $M\overrightarrow{w_1} + H\overrightarrow{w_2}$	5.9	0.048	19.4	0.95
Modifier constituent $\overrightarrow{w_1}$	11.5	0.003	318.3	0.99
Head constituent $\overrightarrow{w_2}$	0	0.932	1	0.5
Additive model $\overrightarrow{w_1} + \overrightarrow{w_2}$	8.0	0.017	55.2	0.98

Table A.1.3: Results of language model comparison in analysis 2.

model	ΔAIC	**Akaike Weight**	**Weight ratio**	**Normalized probability**
CAOSS $M\overrightarrow{w_1} + H\overrightarrow{w_2}$	11.3	0.003	288.8	0.99
Modifier constituent $\overrightarrow{w_1}$	13.0	0.002	656.3	0.99
Head constituent $\overrightarrow{w_2}$	0	0.990	1	0.5
Additive model $\overrightarrow{w_1} + \overrightarrow{w_2}$	10.6	0.005	196.2	0.99

Table A.1.4: Results of language model comparison in analysis 3.

model	ΔAIC	**Akaike Weight**	**Weight ratio**	**Normalized probability**

CAOSS $M\overrightarrow{w_1} + H\overrightarrow{w_2}$	8.1	0.017	57.3	0.98
Modifier constituent $\overrightarrow{w_1}$	14.8	<0.001	1610.3	0.99
Head constituent $\overrightarrow{w_2}$	0	0.979	1	0.5
Additive model $\overrightarrow{w_1} + \overrightarrow{w_2}$	11.5	0.003	310.7	0.99

A.1.6 Reaction times across conditions

The time taken to select an image differed between conditions. According to Welch's t-tests, when removing the slowest and fastest 0.5% of reaction times (i.e., considering reaction times within the 0.5 and 99.5 percentiles), log-transformed responses in the head-compound condition (analysis 3; $M = 0.38, SD = 0.56$) were faster than those of the between-interpretation condition (analysis 1; $M = 0.42, SD = 0.54$), $t(6509.7) = -3.00, p = 0.0027$, and of the within-interpretation condition (analysis 2; $M = 0.54, SD = 0.56$), $t(6240.5) = -13.32, p < 0.0001$, while responses in the between-interpretation condition were faster than those of the within-interpretation condition, $t(6321.8) = -9.17, p < 0.0001$. These results remain significant ($p < 0.01$) even when considering all reaction time data (i.e., below the 0.5 and above the 99.5 percentiles).

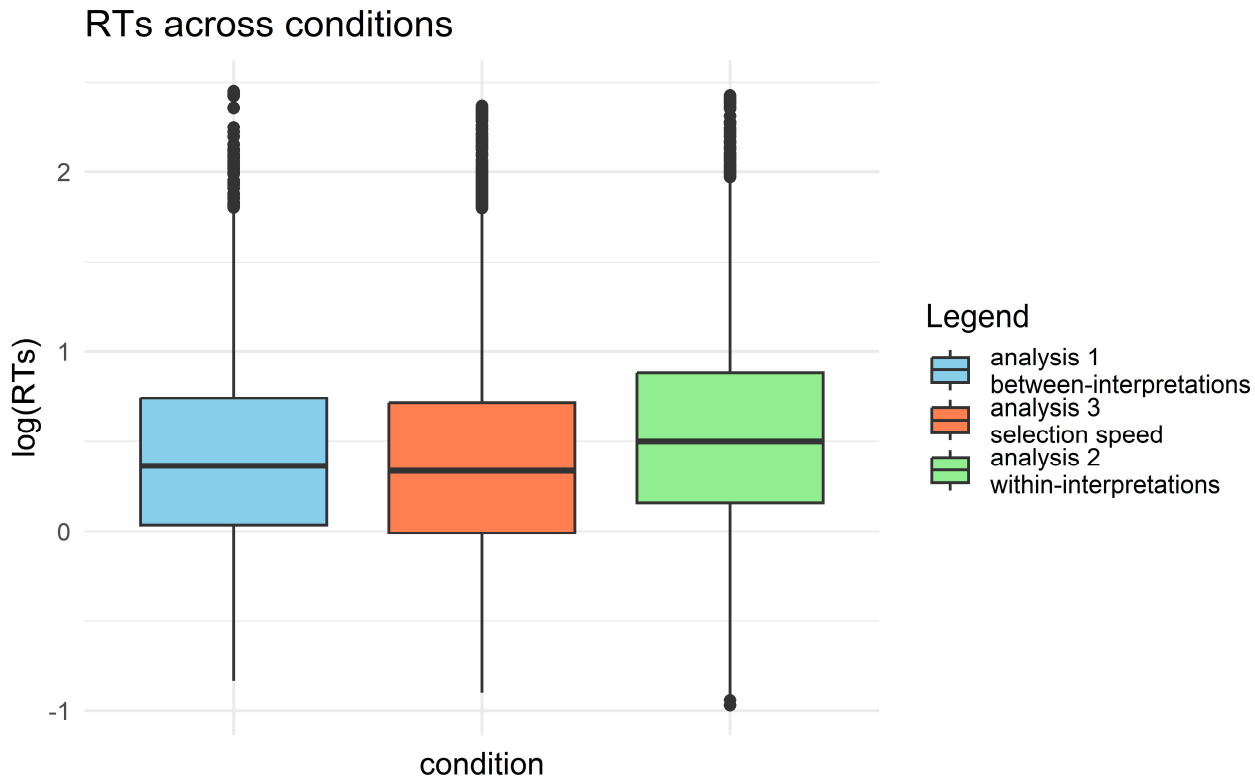

Figure A.1.4: log-transformed rection times across conditions, considering responses within the 0.5 and 99.5 percentiles.

Appendix A.2

A.2.1 Tables of results

Table A.2.1: Core ROIs results. Right-tailed Wilcoxon singed-rank test results for all models. Test statistics were computed over the Spearman correlation coefficients (i.e., subject-level RSA results) for all subjects and all studies (N = 85). RSA were performed with Spearman correlations; partial RSA controlling for confounding models were performed with partial Spearman correlations (models controlled in such manner are specified with the *p* subscript). The table reports Z scores and uncorrected *p*-values; results significant after Bonferroni correction (adjusted $\alpha = .0125$) are reported in **bold.**

RDM predictor	LATL		LAG		RAG		LIFG	
	Z	p	Z	p	Z	p	Z	p
run	**7.186**	**<.001**	**7.808**	**<.001**	**7.554**	**<.001**	**7.195**	**<.001**
$Levenshtein\ distance$	-0.272	.607	-0.911	.819	0.438	.331	-0.311	.622
$w_1\ frequency$	-0.583	.720	-0.074	.530	-0.245	.597	-0.193	.576
$w_2\ frequency$	-0.692	.756	0.232	.408	1.726	.042	-0.596	.724

$concreteness$	0.355	.361	0.079	.469	-0.197	.578	0.241	.405
w_{1_p}	1.503	.066	1.209	.113	0.508	.306	*2.138*	*.016*
w_{2_p}	0.618	.268	**2.892**	**.002**	0.140	.444	1.481	.069
$additive_p$	1.393	.082	**3.558**	**<.001**	1.315	.094	1.998	.023
$multiplicative_p$	*2.090*	*.018*	1.323	.093	0.539	.295	**2.300**	**.011**
$CAOSS_p$	0.206	.418	0.500	.309	0.364	.358	0.018	.493

Table A.2.2: Peak activations for the searchlight RSA in the semantic network. Coordinates [x, y, z] in Montreal Neurological Institute (MNI) space and selection of cluster maxima according to SPM12. Searchlight RSA were performed with partial Spearman correlations controlling for confounding models. All results displayed here were obtained controlling for baseline models.

model	**Cluster *p* (FWE)**	**Cluster size**	**Peak Z**	**MNI**		
$multiplicative$	.0004	452	3.74	-64	-12	2
			3.60	-60	-22	2
			3.42	-66	-24	-8
	.030	186	3.55	-58	24	8
			3.47	-44	26	8
			3.20	-52	26	0

Table A.2.3: Peak activations for the searchlight RSA in the semantic network. Coordinates [x, y, z] in Montreal Neurological Institute (MNI) space and selection of cluster maxima according to SPM12. Searchlight RSA were performed with partial Spearman correlations controlling for confounding models. All results displayed here were obtained controlling for baseline models and non-compositional semantic models

model	**Cluster *p* (FWE)**	**Cluster size**	**Peak Z**	**MNI**		
$multiplictive$	.0087	246	3.73	-64	-12	2
			3.50	-60	-22	4
			3.46	-66	-24	-8

Appendix A.3

A.3.1 By-layer model results

By-layer results of all LLMs and baseline models employed are reported below, including both model coefficients from negative binomial models and AIC values.

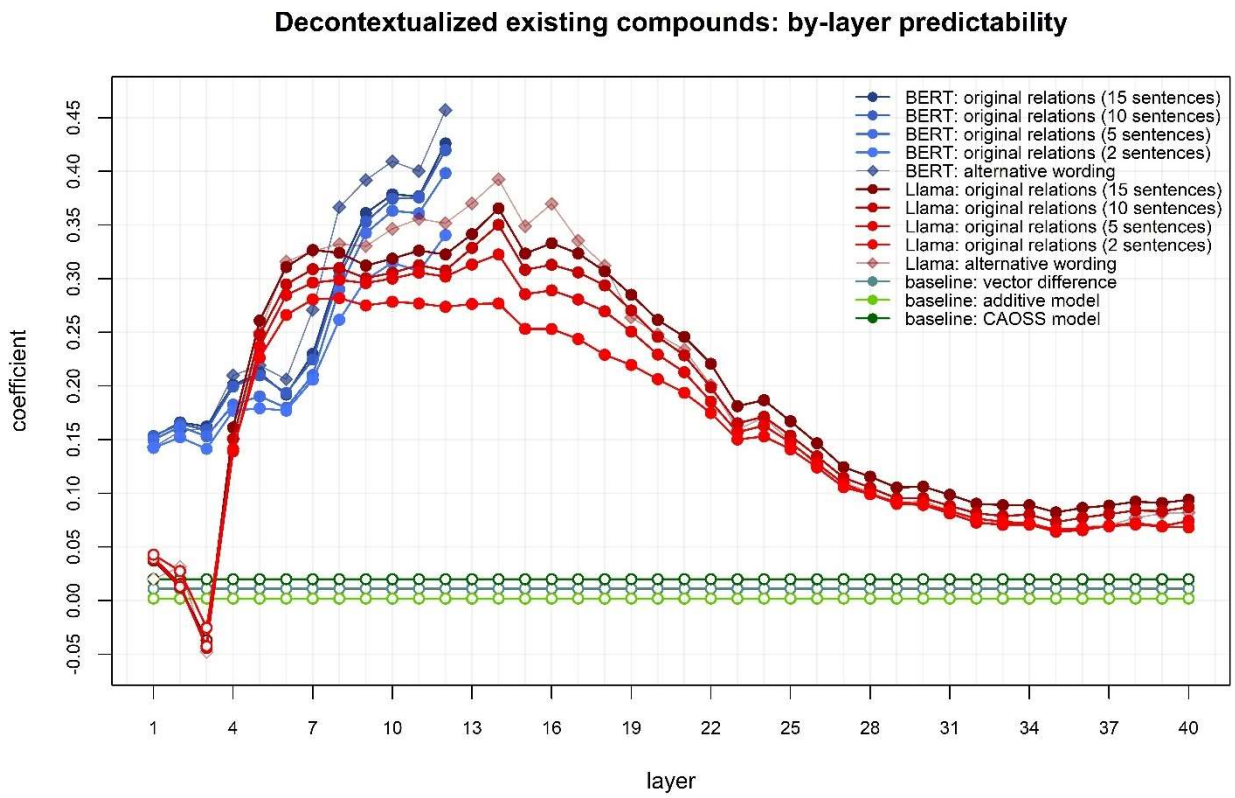

Figure A.3.1: Negative binomial coefficients for the main effect of relation probability on (count) selection frequency. Filled symbols indicate that the corresponding effect is significant after Bonferroni correction for the number of layers.

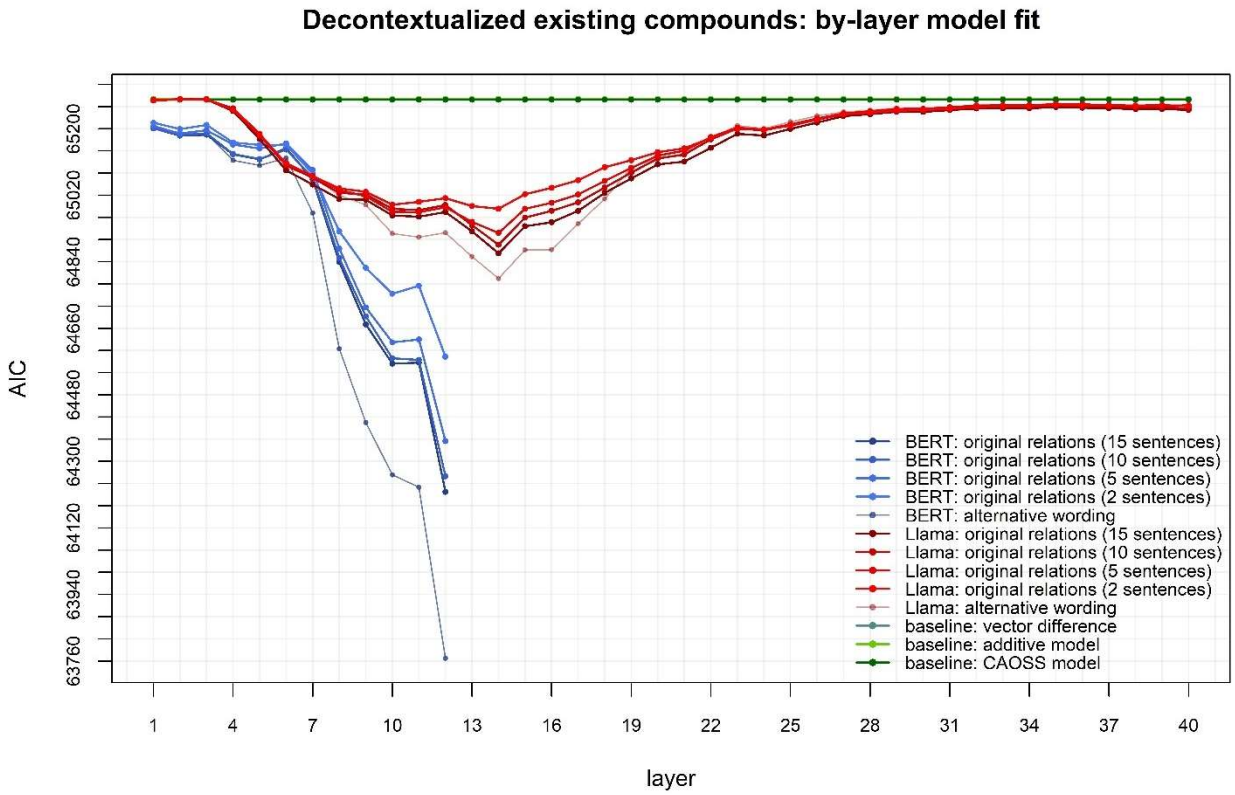

Figure A.3.2: AIC values of the generalized linear mixed models with relation probability as the main predictor.

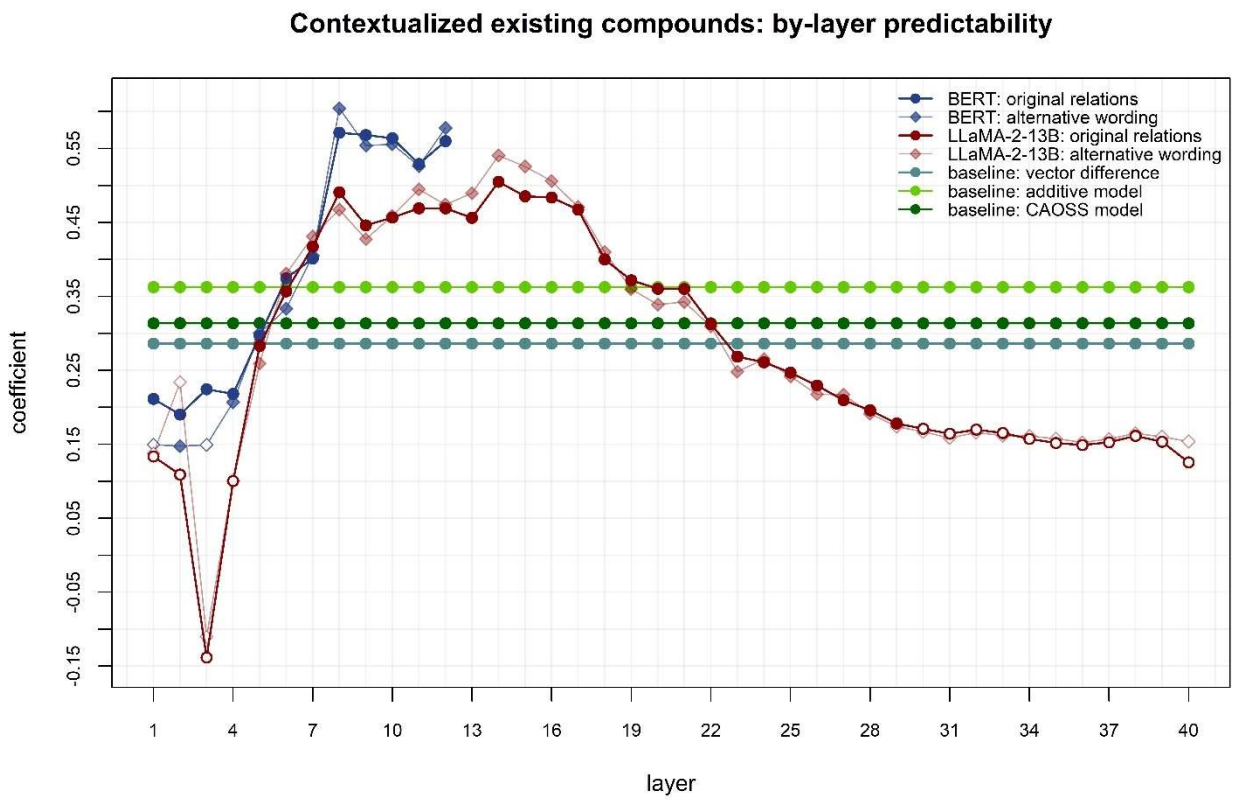

Figure A.3.3: Negative binomial coefficients for the main effect of relation probability on (count) selection frequency. Filled symbols indicate that the corresponding effect is significant after Bonferroni correction for the number of layers.

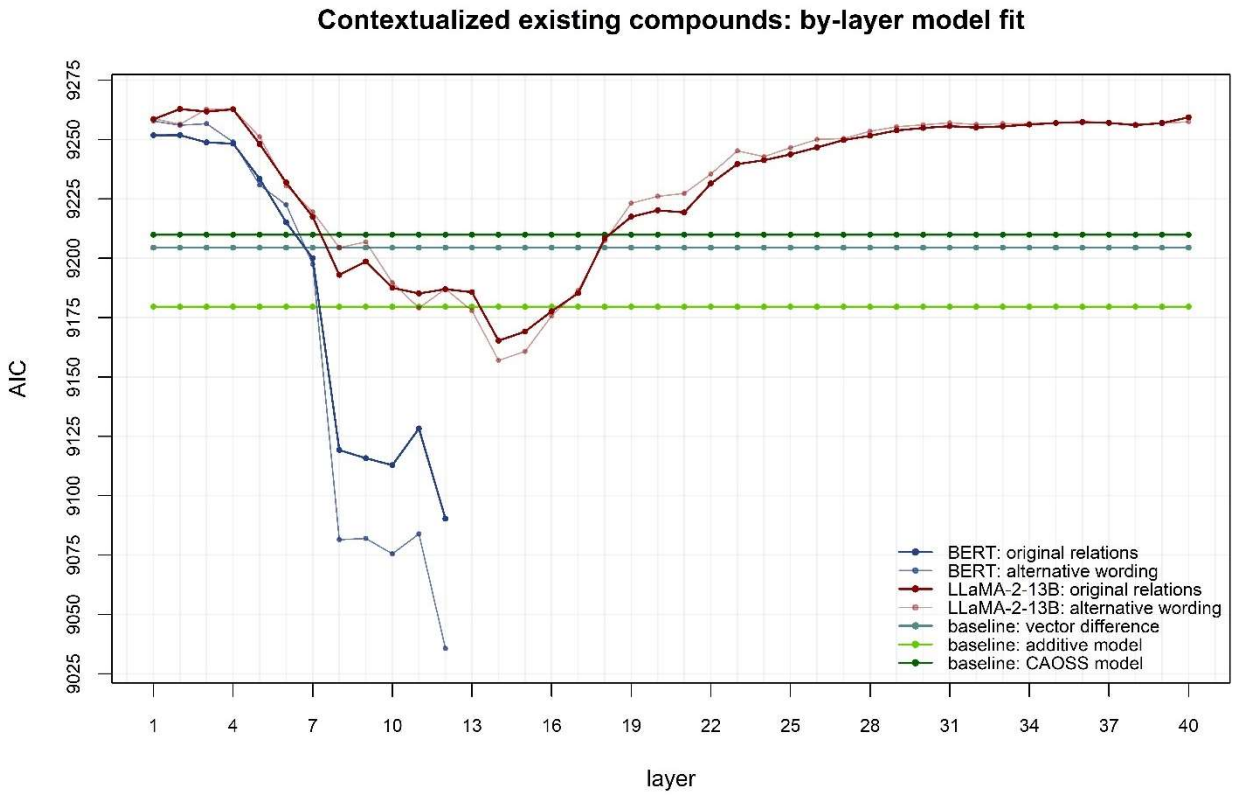

Figure A.3.4: AIC values of the generalized linear mixed models with relation probability as the main predictor.

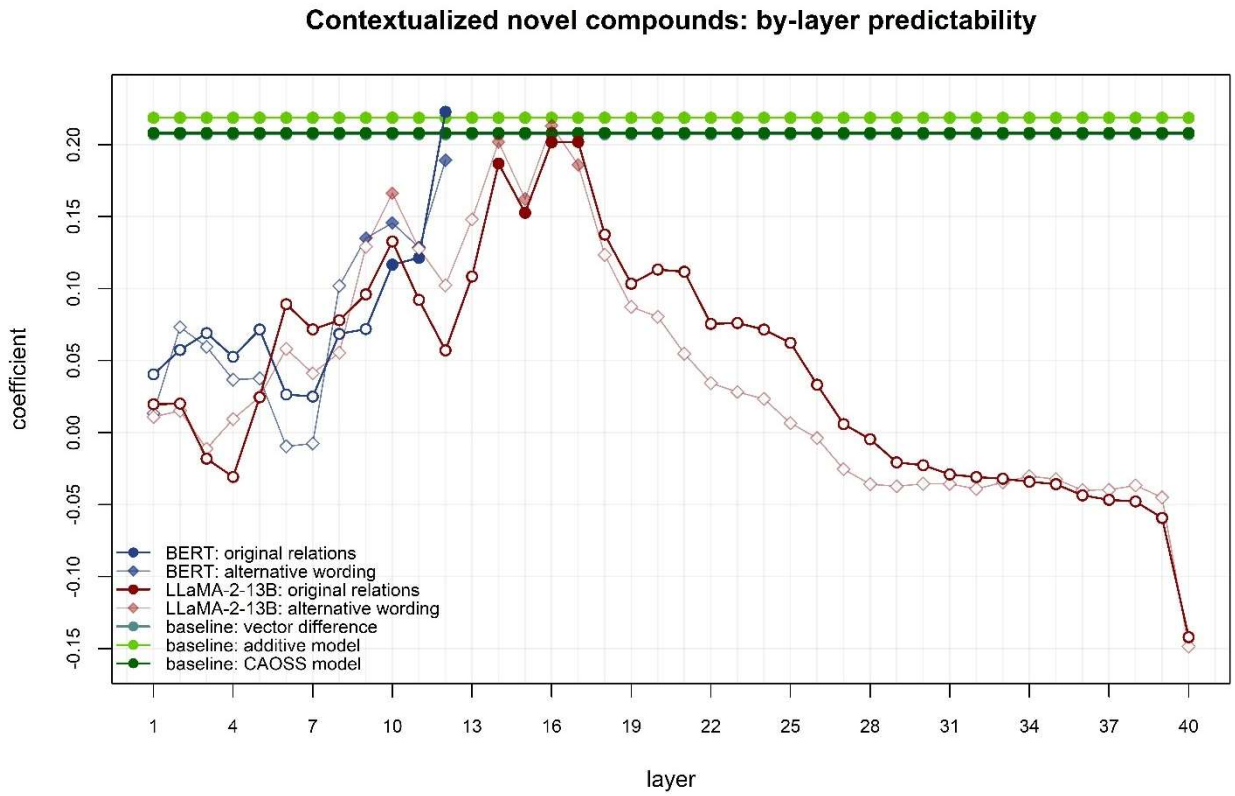

Figure A.3.5: Negative binomial coefficients for the main effect of relation probability on (count) selection frequency. Filled symbols indicate that the corresponding effect is significant after Bonferroni correction for the number of layers.

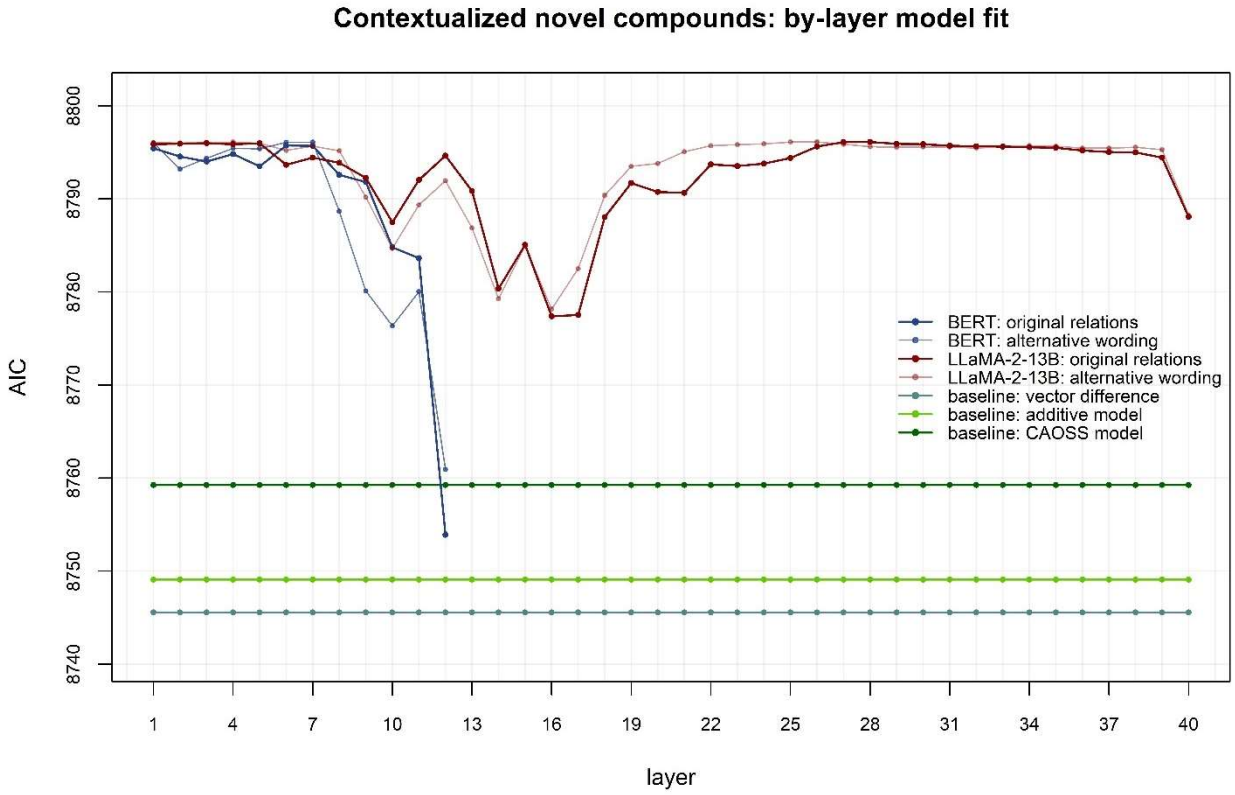

Figure A.3.6: AIC values of the generalized linear mixed models with relation probability as the main predictor.

A.3.2: Statistical models assumptions

Negative Binomial model assumptions were assessed with the R package DHARMa (Hartig, 2020). Zero-inflation tests over unconditional simulations showed no zero-inflation (values ranged from 0.99 to 1.01 across all layers of BERT-base and Llama-2-13b, and across all baseline models). Across all studies, dispersion tests over conditional simulations revealed small to moderate underdispersion, suggestive of overfitting and of overly conservative *p*-values (overdispersion in BERT-base ranged from 0.59 to 0.85 across layers and studies, 0.67 to 0.86 for Llama-2-13b, 0.60 to 0.86 for the baseline models). The expected and observed residuals showed slight deviations from theoretically ideal patterns. Due to the large number of models evaluated (all layers of all models), this appendix reports only the patterns for BERT-base's 12th layer (figure A.3.7), the best-performing layer and model in analyses 1 and 2, and the third-best in analysis 3.

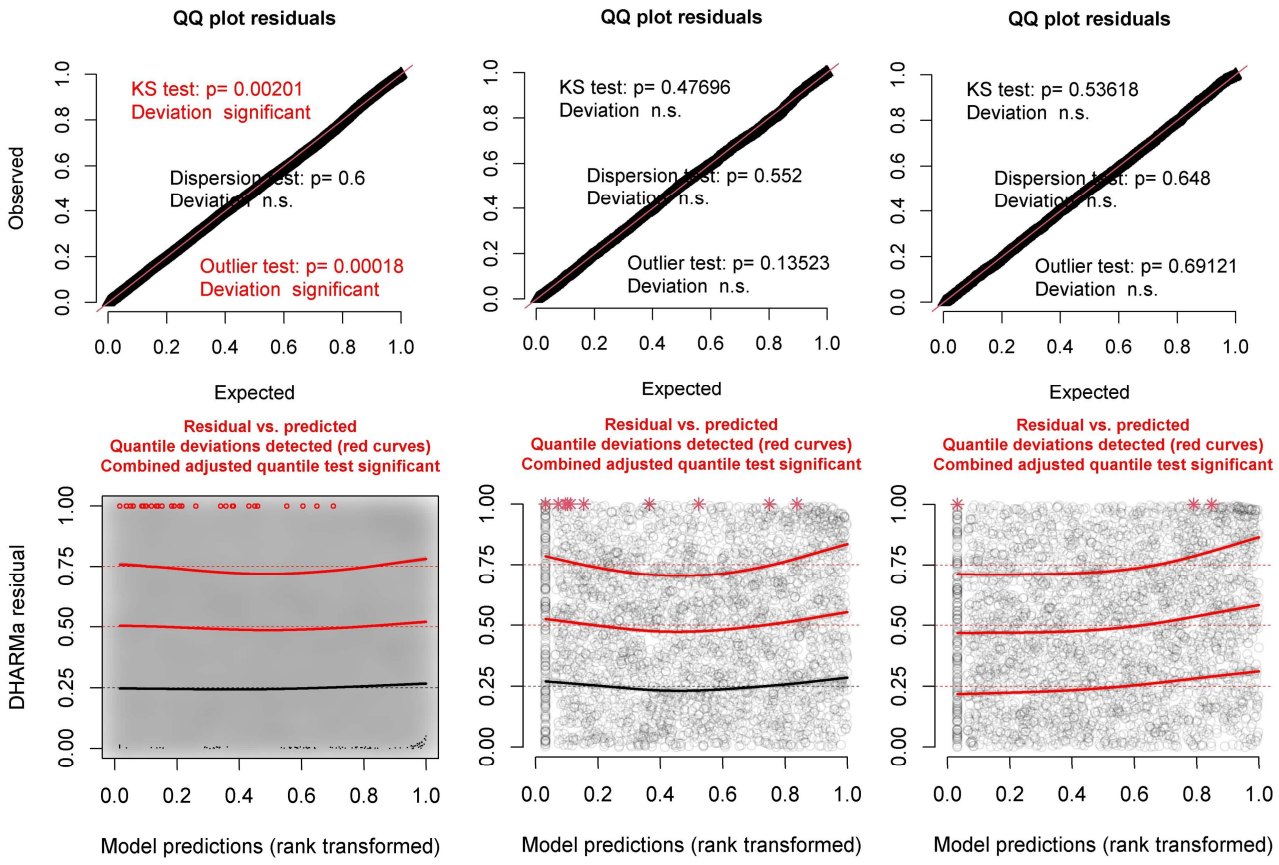

Figure A.3.7*:* Deviations between predicted and observed residuals. Left, middle, and right columns correspond to study 1, 2, and 3, respectively.

While showing slight deviations from the expected trends, these appear to be not large enough to undermine the statistical models. Moreover, deviation tests are likely to be significant at large sample sizes, which is the case for all studies involved (i.e., N = 17920, 3200, and 3200).

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Resarch Methods*, *40(1)*, 278–289.

Amenta, S., Günther, F., & Marelli, M. (2020). A (distributional) semantic perspective on the processing of morphologically complex words. *The Mental Lexicon*, *15*(1), 62–78.

Auch, L., Gagné, C. L., & Spalding, T. L. (2020). Conceptualizing semantic transparency: A systematic analysis of semantic transparency measures in English compound words. *Methods in Psychology*, *3*(May), 100030.

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). *Frequency effects in compound processing*. 257–270.

Baayen, R. H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words. *The Mental Lexicon*, *2*(3), 419–463.

Baayen, R. Harald, Milin, P., rević, D. F., Hendrix, P., & Marelli, M. (2011). An Amorphous Model for Morphological Processing in Visual Comprehension Based on Naive Discriminative Learning. *Psychological Review*, *118*(3), 438–481.

Baggio, G. (2018). *Meaning in the brain*. MIT Press.

Baggio, G. (2021). Compositionality in a parallel architecture for language processing. *Cognitive Science*, *45*(5), e12949.

Baron, S. G., & Osherson, D. (2011). Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage*, *55*(4), 1847–1852.

Baroni, M. (2013). Composition in distributional semantics. *Linguistics and Language Compass*, *7*(10), 511–522.

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in Space: A Program of Compositional Distributional Semantics. *Linguistic Issues in Language Technology*, *9*(6), 242–346.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The waCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, *1*, 238–247.

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. *EMNLP 2010 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, *October*, 1183–1193.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*(1), 617-645.

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

Barsalou, L. W. (2020). Challenges and Opportunities for Grounding Cognition. *Journal of Cognition*, *3*(1), 1–24.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1).

Bender, E. M., & Koller, A. (2020). *Climbing towards NLU*. *2*, 5185–5198.

Benjamin, S., & Schmidtke, D. (2023). Conceptual combination during novel and existing compound word reading in context: A self-paced reading study. *Memory and Cognition*.

Betker, J., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., & Guo, Y. (2023). DALLE3: Improving Image Generation with Better Captions. *ArXiv*.

Beyersmann, E., Kezilas, Y., Coltheart, M., Castles, A., Ziegler, J. C., Taft, M., & Grainger, J. (2018). Taking the book from the bookshelf: Masked constituent priming effects from compound words and nonwords. *Journal of Cognition*, *1*(1), 1–13.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796.

Bommasani, R., Davis, K., & Cardie, C. (2020). *Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings*. 4758–4781.

Boylan, C., Trueswell, J. C., & Thompson-Schill, S. L. (2017). Relational vs. attributive

interpretation of nominal compounds differentially engages angular gyrus and anterior temporal lobe. *Brain and Language*, *169*, 8–21.

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1).

Bruni, E., & Baroni, M. (2014). *Multimodal Distributional Semantics*. *49*, 1–47.

Brysbaert, M., & Diependaele, K. (2013). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods*, *45*(2), 422–430.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, *51*(4), 1849–1863.

Buijtelaar, L., & Pezzelle, S. (2023). *A Psycholinguistic Analysis of BERT's Representations of Compounds*.

Carota, F., Kriegeskorte, N., Nili, H., & Pulvermüller, F. (2017). Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. *Cerebral Cortex*, *27*(1), 294–309.

Cassani, G., Günther, F., Attanasio, G., Bianchi, F., & Marelli, M. (2023). *Meaning Modulations and Stability in Large Language Models: An Analysis of BERT Embeddings for Psycholinguistic Research*.

Chamberlain, J. M., Gagné, C. L., Spalding, T. L., & Lõo, K. (2019). Detecting Spelling Errors in Compound and Pseudocompound Words. *Journal of Experimental Psychology: Learning Memory and Cognition*, *46*(3), 580–602.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). *Return of the Devil in the Details: Delving Deep into Convolutional Nets*. 137–150.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22-29.

Cohen, B., & Murphy, G. L. (1984). *Models of Concepts*. *58*, 27–58.

Coil, J., & Shwartz, V. (2023). *From chocolate bunny to chocolate crocodile: Do Language Models Understand Noun Compounds? 3*(2018).

Costello, F. J., & Keane, M. T. (2000). Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science*, *24*(2), 299–349.

Coutanche, M. N., Solomon, S. H., & Thompson-Schill, S. L. (2018). Conceptual Combination. *The Big Book of Concepts*, 1–22.

Coutanche, M. N., & Thompson-schill, S. L. (2015). *Creating Concepts from Converging Features in Human Cortex*. *September*, 2584–2593.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, *23*(3), 371-414.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Resarch Methods*, *51*(3), 987–1006.

del Prado Martín, F. M., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H., & Baayen, R. H. (2005). Changing places: A cross-language perspective on frequency and family size in Dutch and Hebrew. *Journal of Memory and Language*, *53*(4), 496-512.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186.

Dinu, G., Pham, N. T., & Baroni, M. (2013a). DISSECT - DIS tributional SE mantics C omposition T oolkit. *Proceedings of the System Demonstrations of ACL*, *2010*, 31–36.

Dinu, G., Pham, N. T., & Baroni, M. (2013b). General estimation and evaluation of compositional distributional semantic models. *Proceedings of the Workshop on Continuous Vector Space Models and Their Compositionality*, 50–58.

Downing, P. (1977). *On the Creation and Use of English Compound*. *53*(4), 810–842.

El-Bialy, R., Gagné, C. L., & Spalding, T. L. (2013). Processing of English compounds is sensitive to the constituents' semantic transparency. *The Mental Lexicon*, *8*(1), 75–95.

Estes, Z., & Glucksberg, S. A. M. (2000). *Interactive property attribution in concept combination*. *28*(1), 28–34.

Evert, S. (2015). *Some theoretical and experimental observerations on na ¨ıve discriminative learning*.

Falkauskas, K., & Kuperman, V. (2015). When experience meets language statistics: Individual variability in processing English compound words. *Journal of Experimental Psychology: Learning Memory and Cognition*.

Forgács, B., Bohrn, I., Baudewig, J., Hofmann, M. J., Pléh, C., & Jacobs, A. M. (2012). Neural correlates of combinatorial semantic processing of literal and figurative noun noun compound words. *NeuroImage*, *63*(3), 1432–1442.

Frankland, S. M., & Greene, J. D. (2020a). Concepts and Compositionality: In Search of the Brain's Language of Thought. *Annual Review of Psychology*, *71*, 273–303.

Frankland, S. M., & Greene, J. D. (2020b). Two Ways to Build a Thought: Distinct Forms of Compositional Semantic Representation across Brain Regions. *Cerebral Cortex*, *30*(6), 3838–3855.

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.

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning Memory and Cognition*, *23*(1), 71–87.

Gagné, C. L., & Spalding, T. L. (2004). Effect of discourse context and modifier relation frequency on conceptual combination. *Journal of Memory and Language*, *50*(4), 444–455.

Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, *60*(1), 20–35.

Gagné, C. L., & Spalding, T. L. (2013). Conceptual composition: The role of relational competition in the comprehension of modifier-noun phrases and noun-noun compounds. In *Psychology of*

Learning and Motivation - Advances in Research and Theory (Vol. 59).

Gagné, C. L., Spalding, T. L., & Schmidtke, D. (2019). LADEC: The Large Database of English Compounds. *Behavior Research Methods*, *51*(5), 2152–2179.

Gallistel, C. R. (2003). Conditioning from an information processing perspective. *Behavioural Processes*, *62*(1–3), 89–101.

Gatti, D., Marelli, M., & Rinaldi, L. (2023). Out-of-vocabulary but not meaningless: Evidence for semantic-priming effects in pseudoword processing. *Journal of Experimental Psychology: General*, *152*(3), 851.

Gatti, D., Marelli, M., Vecchi, T., & Rinaldi, L. (2022). Spatial Representations Without Spatial Computations. *Psychological Science*, *33*(11), 1947–1958.

Glenberg, A. M., Witt, J. K., & Metcalfe, J. (2013). *From the Revolution to Embodiment : 25 Years of Cognitive Psychology*.

Graessner, A., Zaccarella, E., Friederici, A. D., Obrig, H., & Hartwigsen, G. (2021). Dissociable contributions of frontal and temporal brain regions to basic semantic composition. *Brain Communications*, *3*(2).

Graessner, A., Zaccarella, E., & Hartwigsen, G. (2021). Differential contributions of left-hemispheric language regions to basic semantic composition. *Brain Structure and Function*, *226*(2), 501–518.

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.

Graves, W. W., Binder, J. R., Desai, R. H., Conant, L. L., & Seidenberg, M. S. (2010). Neural correlates of implicit and explicit combinatorial semantic processing. *NeuroImage*, *53*(2), 638–646.

Graves, W. W., Levinson, H., Coulanges, L., Cahalan, S., Cruz, D., Sancimino, C., Bal, V. H., & Rosenberg-Lee, M. (2022). Neural differences in social and figurative language processing on the autism spectrum. *Neuropsychologia*, *171*(April), 108240.

Greco, C. M., B, A. T., & Zumpano, E. (2022). *A Comparison of Transformer-Based Language Models on NLP Benchmarks*. 490–501.

Green, J. A. (2021). Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. *Health Psychology and Behavioral Medicine*, *9*(1), 436–455.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). *Topics in Semantic Representation*. *114*(2), 211–244.

Guevara, E. (2010). A Regression Model of Adjective-Noun Compositionality in Distributional Semantics. *In Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics, ACL 2010, Uppsala, Sweden, 16 July 2010*, *July*, 33–37. papers3://publication/uuid/604A9434-ABA4-4987-AC71-A998FB8A9B8F

Günther, F., & Marelli, M. (2016). Understanding karma police: The perceived plausibility of noun compounds as predicted by distributional models of semantic representation. *PLoS ONE*, *11*(10), 1–36.

Günther, F., & Marelli, M. (2018). Enter Sandman: Compound Processing and Semantic Transparency in a Compositional Perspective. *Journal of Experimental Psychology: Learning Memory and Cognition*, *45*(10), 1872–1882.

Günther, F., & Marelli, M. (2020). Trying to make it work: Compositional effects in the processing of compound "nonwords." *Quarterly Journal of Experimental Psychology (2006)*, *73*(7), 1082–1091.

Günther, F., & Marelli, M. (2021a). CAOSS and transcendence : Modeling role-dependent constituent meanings in compounds. *Morphology*.

Günther, F., & Marelli, M. (2021b). *Patterns in CAOSS: Distributed representations predict variation in relational interpretations for familiar and novel compound words*. *46*(1), 46–65.

Günther, F., & Marelli, M. (2022). Patterns in CAOSS : Distributed representations predict variation in relational interpretations for familiar and novel compound words. *Cognitive Psychology*, *134*(January), 101471.

Günther, F., Marelli, M., & Bölte, J. (2020). Semantic transparency effects in German compounds: A large dataset and multiple-task investigation. *Behavior Research Methods*, *52*(3), 1208–1224.

Günther, F., Marelli, M., & Petilli, M. A. (2023). *The challenge of representation learning: Improved accuracy in deep vision models does not come with better predictions of perceptual*

similarity.

Günther, F., Marelli, M., Tureski, S., & Petilli, M. (2022). ViSpa (Vision Spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. In *Psychological Review*.

Günther, F., Petilli, M. A., & Marelli, M. (2020). Semantic transparency is not invisibility: A computational model of perceptually-grounded conceptual combination in word processing. *Journal of Memory and Language*, *112*(July 2019), 104104.

Günther, F., Petilli, M. A., Vergallito, A., & Marelli, M. (2020). Images of the unseen: extrapolating visual representations for abstract and concrete words in a data-driven computational model. *Psychological Research*, *0123456789*.

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.

Harris, Z. S. (1954). Distributional Structure. WORD, *10*(2–3), 146–162.

Hartig, F. (2020). DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. *R Package*, *1*.

Hoffman, P., McClelland, J. L., & Lambon-Ralph, M. A. (2017). Concepts, control and context. *Psychological Review*, *125*(3), 293–328.

Hollis, G. (2017). Estimating the average need of semantic knowledge from distributional semantic models. *Memory and Cognition*, *45*(8), 1350–1370.

Hoppe, D. B., Hendriks, P., Ramscar, M., & van Rij, J. (2022). An exploration of error-driven learning in simple two-layer networks from a discriminative learning perspective. *Behavior Research Methods*, *54*(5), 2221-2251.

David Hume, & Peter Millican. (2007). An Enquiry Concerning Human Understanding. OUP Oxford.

Humphreys, G. F., Lambon Ralph, M. A., & Simons, J. S. (2021). A Unifying Account of Angular Gyrus Contributions to Episodic and Semantic Cognition. *Trends in Neurosciences*, *44*(6), 452–463.

Huth, A. G., Heer, W. A. De, Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural

speech reveals the semantic maps that tile human cerebral cortex. *Nature*.

Jackson, R. L. (2021). The neural correlates of semantic control revisited. *NeuroImage*, *224*(July 2020), 117444.

Ji, H., Gagné, C. L., & Spalding, T. L. (2011). Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds. *Journal of Memory and Language*, *65*(4), 406–430.

Jones, L. L., Estes, Z., & Marsh, R. L. (2008). *An asymmetric effect of relational integration on recognition memory*. *61*(8), 1169–1176.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review*, *122*(3), 570–574.

Jones, M. N., & Mewhort, D. J. K. (2007). *Representing Word Meaning and Order Information in a Composite Holographic Lexicon*. *114*(1), 1–37.

Juhasz, B. J., Lai, Y. H., & Woodcock, M. L. (2014). A database of 629 English compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods*, *47*(4), 1004–1019.

Kenett, Y. N., & Thompson-Schill, S. L. (2020). Novel conceptual combination can dynamically reconfigure semantic memory networks. *The Japanese Journal of Rehabilitation Medicine*, *34*(3), 234–235.

Kilgarriff, A., Mcadam, K., Rundell, M., & Rychlý, P. (2007). GDEX : Automatically Finding Good Dictionary Examples in a Corpus 2 . A learner's dictionary project. *Eura*, *Kilgarriff 2006*, 425–432.

Kim, S. Y., Yap, M. J., & Goh, W. D. (2019). The role of semantic transparency in visual word recognition of compound words: A megastudy approach. *Behavior Research Methods*, *51*(6), 2722–2732.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(NOV), 1–28.

Kumar, A. A. (2020). Semantic memory: A review of methods, models, and current challenges. In *Psychonomic Bulletin and Review*. Psychonomic Bulletin & Review.

Kumar, A. A., Steyvers, M., & Balota, D. A. (2021). A Critical Review of Network-Based and Distributional Approaches to Semantic Memory Structure and Processes. *Topics in Cognitive Science*, *June*.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading Polymorphemic Dutch Compounds: Toward a Multiple Route Model of Lexical Processing. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 876–895.

Lake, B. M., & Murphy, G. L. (2021). Word Meaning in Minds and Machines. *Psychological Review*.

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 of the United States of America*, *107*(6), 2717–2722.

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.

Levy, O., & Goldberg, Y. (2014, June). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning* (pp. 171-180).

Li, S., Carlson, R., & Potts, C. (2022). Systematicity in GPT-3's Interpretation of Novel English Noun Compounds. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 717–728.

Libben, G. (2010). Why Study Compound Processing? An overview of the issues. *The Representation and Processing of Compound Words*, 1–22.

Libben, G. (2014). The nature of compounds: A psychocentric perspective. *Cognitive Neuropsychology*, *31*(1–2), 8–25.

Libben, G. (2017). The quantum metaphor and the organization of words in the mind. *Journal of Cultural Cognitive Science*, *1*(1), 49–55.

Libben, G., Derwing, B. L., & De Almeida, R. G. (1999). Ambiguous novel compounds and models of morphological parsing. *Brain and Language*, *68*(1–2), 378–386.

Libben, G., Gagné, C. L., & Dressler, W. U. (2020). The Representation and Processing of Compound Words. *Word Knowledge and Word Usage*, 336.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031.

Liuzzi, G. A., Ubaldi, S., & Laurence, S. (2021). Representations of conceptual information during automatic and active semantic access. *Neuropsychologia*, *160*(July), 107953.

Logie, R. (2018). Human Cognition: Common Principles and Individual Variation. *Journal of Applied Research in Memory and Cognition*, *7*(4), 471–486.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302.

Louwerse, M. M., & Benesh, N. (2012). Representing spatial structure through maps and language: Lord of the rings encodes the spatial structure of middle earth. *Cognitive Science*, *36*(8), 1556–1569.

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.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, *28*(2), 203–208.

Lynott, D., & Connell, L. (2010). Embodied conceptual combination. *Frontiers in Psychology*, *1*(NOV), 1–14.

Maguire, P., Maguire, R., & Cater, A. W. S. (2010). The Influence of Interactional Semantic Patterns on the Interpretation of Noun-Noun Compounds. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(2), 288–297.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, *122*(3), 485–515.

Marelli, M., Gagné, C. L., & Spalding, T. L. (2017). Compounding as Abstract Operation in Semantic Space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, *166*, 207–224.

Marelli, M., & Luzzatti, C. (2012). Frequency effects in the processing of Italian nominal

compounds: Modulation of headedness and semantic transparency. *Journal of Memory and Language*, *66*(4), 644–664.

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X., Petrides, M., Jefferies, E., & Smallwood, J. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(44), 12574–12579.

Markman, A. B., & Rein, J. R. (2013). The Nature of Mental Concepts. *The Oxford Handbook of Cognitive Psychology*, *July 2016*, 1–11.

Marr, D. (1982). Vision: A computational approach. Freeman & Co.

Martin, D. I., & Barry, M. W. (2007). *Mathematical Foundations Behind Latent Semantic Analysis*. 35–55.

McRae, K. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments, & Computers*, *40*(1), 183–190.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. Psychological review, 85(3), 207.

Mikolov, T. (2013). Distributed Representations of Words and Phrases and their Compositionality. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1389–1399.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2nd Workshop on Computational Linguistics for Literature, CLfL 2013 at the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2013*, 746–751.

Miletic, F., & Schulte, S. (2023). *A Systematic Search for Compound Semantics in Pretrained BERT Architectures*. 1499–1512.

Misra, K. (2022). *minicons: Enabling Flexible Behavioral and Representational Analyses of*

Transformer Language Models.

Mitchell, J., & Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, *34*(8), 1388–1429.

Mitchell, T. M., Malave, L., Shinkareva, V., & Mason, A. (2008). *Predicting Human Brain Activity Associated with the Meanings of Nouns*. *320*(May), 1191–1195.

Molinaro, N., Paz-Alonso, P. M., Duñabeitia, J. A., & Carreiras, M. (2015). Combinatorial semantics strengthens angular-anterior temporal coupling. *Cortex*, *65*, 113–127.

Momenian, M., Radman, N., Rafipoor, H., Barzegar, M., & Weekes, B. (2021). Compound words are decomposed regardless of semantic transparency and grammatical class: An fMRI study in Persian. *Lingua*, *259*.

Moore, M., & Gordon, P. C. (2014). Reading ability and print exposure: item response theory analysis of the author recognition test. *Behavior Research Methods*, *47*(4), 1095–1109.

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, *103*, 130–138.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643.

Murphy, G. L. (1988). *Comprehending Complex Concepts*. *562*, 529–562.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). *Evaluating Vector-Space Models of Word Representation , or , The Unreasonable Effectiveness of Counting Words Near Other Words*.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4).

Ormerod, M., Devereux, B., & Martínez del Rincón, J. (2024). How is a “Kitchen Chair” like a “Farm Horse”? Exploring the Representation of Noun-Noun Compound Semantics in Transformer-based Language Models. *Computational Linguistics*, 1–33.

Parrish, A., & Pylkkänen, L. (2022). Conceptual Combination in the LATL With and Without Syntactic Composition. *Neurobiology of Language*, *3*(1), 46–66.

Pavlick, E. (2022). Semantic Structure in Deep Learning. *Annual Review of Linguistics*, *8*, 447–471.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1).

Perfetti, C. (2007). Reading Ability: Lexical Quality to Comprehension. In *Scientific Studies of Reading* (Vol. 11, Issue 4, pp. 375–383).

Perfetti, C. A., & Hart, L. (2002). *The lexical quality hypothesis* (pp. 189–213).

Petrou, M., Tabacchi, M. E., & Piroddi, R. (2010). Networks of concepts and ideas. *Computer Journal*, *53*(10), 1738–1751.

Pezzelle, S., & Marelli, M. (2020). *Do semantic features capture a syntactic classification of compounds? Insights from compositional distributional semantics*.

Pham, H., & Baayen, R. H. (2013). Semantic relations and compound transparency: A regression study in CARIN theory. *Psihologija*, *46*(4), 455–478.

Piantadosi, S. T., & Hill, F. (2022). *Meaning without reference in large language models*.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). *The communicative function of ambiguity in language*. 0–21.

Poortman, E. B., & Pylkkänen, L. (2016). Adjective conjunction as a window into the LATL's contribution to conceptual combination. *Brain and Language*, *160*, 50–60.

Price, Amy R., Bonner, M. F., Peelle, J. E., & Grossman, M. (2015). Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *Journal of Neuroscience*, *35*(7), 3276–3284.

Price, Amy Rose, Peelle, J. E., Bonner, M. F., Grossman, M., & Hamilton, R. H. (2016). Causal evidence for a mechanism of semantic integration in the angular Gyrus as revealed by high-definition transcranial direct current stimulation. *Journal of Neuroscience*, *36*(13), 3829–3838.

Prinz, J. (2004). *Furnishing the Mind: Concepts and Their Perceptual Basis*. MIT Press.

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, *366*(6461), 62–66.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral science*, *12*(5), 410-430.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A.,

Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of Machine Learning Research*, *139*, 8748–8763.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2016). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2021). Hopfield Networks Is All You Need. *ICLR 2021 - 9th International Conference on Learning Representations*.

Ramscar, M. (2019). *Source codes in human communication*. http://arxiv.org/abs/1904.03991

Ran, B., & Duimering, P. R. (2009). *Conceptual Combination : Models, Theories, and Controversies*. *1*(1), 1–9.

Rescorla, R., & Wagner, A. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In *Appleton-Century-Crofts*.

Reverberi, C., Görgen, K., & Haynes, J. D. (2012). Compositionality of rule representations in human prefrontal cortex. *Cerebral cortex*, *22*(6), 1237-1246.

Rodd, J. M. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science*, *15*(2), 411–427.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573-605.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *166*(23), 11537–11546.

Scalise, S., & Bisetto, A. (2005). The Classification of Compounds. *Lingue e Linguaggio*, *4(2)*, *319-*.

Schell, M., Friederici, A. D., & Zaccarella, E. (2022). *Neural classification maps for distinct word combinations in Broca's area*.

Schell, M., Zaccarella, E., & Friederici, A. D. (2017). Differential cortical contribution of syntax and semantics: An fMRI study on two-word phrasal processing. *Cortex*, *96*, 105–120.

Schmidtke, D., Gagné, C. L., Kuperman, V., & Spalding, T. L. (2018). Language experience shapes relational knowledge of compound words. *Psychonomic Bulletin and Review*, *25*(4), 1468–1487.

Schmidtke, D., Gagné, C. L., Kuperman, V., & Tucker, B. V. (2018). *Conceptual relations compete during auditory and visual compound word recognition*. *33*(7), 923–942.

Schmidtke, D., Kuperman, V., Gagné, C. L., & Spalding, T. L. (2016). Competition between conceptual relations affects compound recognition: the role of entropy. *Psychonomic Bulletin and Review*, *23*(2), 556–570.

Schmidtke, D., Van Dyke, A. J., & Kuperman, V. (2018). Individual variability in the semantic processing of English compound words. *Journal of Experimental Psychology: Learning Memory Cognition*, *s9-IX*(217), 146.

Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2021). CompLex: an eye-movement database of compound word reading in English. *Behavior Research Methods*, *53*(1), 59–77.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., Mirman, D., & Coslett, H. B. (2011). *Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain*.

Semenza, C., & Mondini, S. (2010). *Compound Words in Neuropsychology*. *February*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Siew, C. S. Q., Wulff, D. U., Beckage, N. M., & Kenett, Y. N. (2019). Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, *2019*.

Sigman, M., & Cecchi, G. A. (2002). Global organization of the Wordnet lexicon. Proceedings of

the National Academy of Sciences, 99(3), 1742-1747.

Smith, E. E., & Medin, D. L. (1981). Categories and concepts. Harvard University Press.

Soler, A. G., & Apidianaki, M. (2021). Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, *9*, 825–844.

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.

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(43), 1–12.

Spalding, T. L., Gagné, C. L., Mullaly, A., & Ji, H. (2010). Relation-based interpretation of noun-noun phrases: A new theoretical approach. *New impulses in word-formation*, *171*, 283-315.

Spalding, T. L., & Gagné, C. L. (2014). Relational diversity affects ease of processing even for opaque English compounds. *The Mental Lexicon*, *9*(1), 48–66.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., … Schuster, T. (n.d.). *Beyond the Imitation Game : Quantifying and extrapolating the capabilities of language models*.

Stevens, P., & Plaut, D. C. (2022). *From decomposition to distributed theories of morphological processing in reading*.

Stewart, T., & Eliasmith, C. (2011). Compositionality and biologically plausible models. *Oxford Handbook of Compositionality*.

Tang, J., Du, M., Vo, V., Lal, V., & Huth, A. G. (2023). *Brain encoding models based on multimodal transformers can transfer across language and vision*. *NeurIPS*.

Touvron, H., Martin, L., & Stone, K. (2023). *Llama 2 : Open Foundation and Fine-Tuned Chat Models*. *arXiv prep*.

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*, *2017-Decem*(Nips), 5999–6009.

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.

Vig, J. (2019). BertViz: A tool for visualizing multihead self-attention in the BERT model. *International Conference on Learning Representation*, 1–6.

Vigliocco, G., & Vinson, D. P. (2012). Semantic representation. In *The Oxford Handbook of Psycholinguistics* (Issue August).

Vinson, D., & Vigliocco, G. (2008). Feature norms for a large set of object and event concepts. *Behavior Research Methods*, *40*(1), 183–190.

Visser, M., Embleton, K. V, Jefferies, E., Parker, G. J., & Ralph, M. A. L. (2010). *The inferior, anterior temporal lobes and semantic memory clarified: Novel evidence from distortion-corrected fMRI*.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic bulletin & review*, *11*, 192-196.

Walbrin, J., Sossounov, N., Mahdiani, M., Vaz, I., & Almeida, J. (2024). ll Fine-grained knowledge about manipulable objects is well-predicted by contrastive language. *ISCIENCE*, *27*(7), 110297.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, *32*(July), 1–15.

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, *5*(December), 1415–1426.

Wang, W., Hsieh, L., Swamy, G., & Bunge, S. A. (2020). *Transient Neural Activation of Abstract Relations on an Incidental Analogy Task*. 77–88.

Wei, J., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models*. 1–30.

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., & Zemla, J. (2017). R package “corrplot”: Visualization of a Correlation Matrix. *Statistician*, *56*, 316–324.

Westbury, C. (2016). Pay no attention to that man behind the curtain. *The Mental Lexicon*, *11*(3), 350–374.

Westerlund, M., & Pylkkänen, L. (2014). The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia*, *57*(1), 59–70.

Wichmann, F. A., & Geirhos, R. (2023). *Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception ?* 501–524.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. Von, Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. Le, Gugger, S., Drame, M., … Rush, A. M. (2020). *Transformers : State-of-the-Art Natural Language Processing*. 38–45.

Wu, L. ling, & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, *132*(2), 173–189.

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic bulletin & review*, *23*, 1015-1027.

Yu, L., & Ettinger, A. (2020). Assessing phrasal representation and composition in transformers. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 4896–4907.

Yu, L., & Ettinger, A. (2021). On the Interplay Between Fine-tuning and Composition in Transformers. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2279–2293.

Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives without imagery e Congenital aphantasia. *Cortex*, *73*, 378–380.

Zeman, A., Milton, F., Della Sala, S., Dewar, M., Frayling, T., Gaddum, J., Hattersley, A., Heuerman-williamson, B., Jones, K., Mackisack, M., & Winlove, C. (2020). The psychological significance of lifelong visual imagery vividness extremes. *Cortex*, *130*, 426–440.

Zhang, L., & Pylkkänen, L. (2015). NeuroImage The interplay of composition and concept speci fi city in the left anterior temporal lobe : An MEG study. *NeuroImage*, *111*, 228–240.

Zhang, W., Xiang, M., & Wang, S. (2022). The role of left angular gyrus in the representation of linguistic composition relations. *Human Brain Mapping*, *43*(7), 2204–2217.

Zhang, Y., Han, K., Worth, R., & Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, *11*(1), 1–13.

Ziegler, J., & Pylkkänen, L. (2016). Neuropsychologia Scalar adjectives and the temporal unfolding of semantic composition : An MEG investigation. *Neuropsychologia*, *89*, 161–171.

Zipf, G. (1949). Human behavior and the principle of least effort. New York: Addison-Wesley