Two Models of Meaning: Revisiting the Principle of Compositionality from the Neurocognition of Language

RUNNING HEAD: Two Models of Meaning

Noortje J. Venhuizen[1*], Harm Brouwer[1*]

[1]Department of Cognitive Science and Artificial Intelligence, Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands

[*]Correspondence should be addressed to: Noortje Venhuizen (n.j.venhuizen@tilburguniversity.edu) or Harm Brouwer (h.brouwer@tilburguniversity.edu)

Keywords: language comprehension, lexical semantics, utterance meaning, compositionality, Retrieval-Integration theory

Manuscript accepted for publication in *Psychology of Learning and Motivation*

Copyright © 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies. This paper is not the copy of record and may not exactly replicate the authoritative document. The final article is available, upon publication, at: https://doi.org/10.1016/bs.plm.2025.07.007

Abstract

A core tenet in linguistic theory is the *principle of compositionality*, which holds that the meaning of a multi-word utterance directly derives from the meanings of the individual words, and the rules by which they are combined. Semantic theories of lexical word meaning and compositional utterance meaning have, however, developed into surprisingly distinct fields of study. Lexical semantic theories of word meaning focus on modeling conceptual structure and similarity, e.g., the words "tea" and "coffee" are similar in that they both describe drinkable substances. Formal semantic theories focusing on compositional utterance meaning, in turn, focus on modeling sentence- and discourse-level entailments and inferences, e.g., "drinking hot coffee" entails "drinking coffee". Critically, attempts at unifying models of lexical and compositional semantics have proven challenging and often yield complex frameworks, in which word- and utterance-level meanings are patched together to form a whole, without fully integrating their semantic contributions. We here revisit the principle of compositionality from the neurocognition of language, which reveals that the human comprehension system harnesses distinct models for lexical and compositional meaning, and that these models are critically intertwined in a cyclic architecture for language comprehension. Within this architecture, compositionality arises from a non-linear mapping of lexical semantic representations into a space for compositional semantic meaning, resulting in a continuous, expectation-based, and spatiotemporally-extended notion of compositional integration. This novel perspective on compositionality, combining linguistic and neurocognitive theory, paves way for more integrative approach towards modeling the meaning of words and utterances.

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

1 Introduction

One of the core topics in linguistic theory has traditionally been the question of how the meaning of complex multi-word utterances is derived from the meaning of the individual words that constitute these utterances. In the traditional view, there is a clear separation between the *syntactic* principles that determine how words can be combined to form complex utterances, and the *semantic* principles that define how meanings are represented and constructed. This distinction is colorfully illustrated in the famous example "Colorless green ideas sleep furiously", which was introduced as an example of a sentence that is grammatically correct, yet nonsensical (Chomsky, 1957, p.15). This distinction between syntax and semantics has long been a guiding principle in answering the overarching question of how meaning is assigned to linguistic input. Specifically, it has led to the fundamental principle that the meaning of a complex expression is fully determined by the meanings of the individual words that constitute the expression, and the way that they are combined (Partee, 1995). This *principle of compositionality* lies at the core of current approaches in semantic theory, which presuppose a close relationship between the lexical meanings of individual words and the compositional meanings assigned to sentences and utterances; that is, utterance-level meaning is directly derived from the meanings of the individual words and the syntactic rules by which they are combined.

The close formal relationship between lexical and compositional meaning that is assumed by the principle of compositionality has some desirable properties, as it explains the observation that human language users are able to produce and understand an infinitely large number of complex expressions that they have not encountered before (referred to as *productivity* of language use), and that they can systematically combine and reorder the constituents of complex expressions into novel utterances (*systematicity* of language use). While the principle of compositionality takes center stage in explaining

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

these premises of language use, semantic theories that study lexical meaning at the level of words and and those that focus on compositional meaning at the level of sentences and discourses have developed into surprisingly distinct fields of study.

Lexical semantic (LS) theories aim to model the meaning of individual words. In particular, distributional approaches to LS model word meaning as vector representations derived from semantic features, capturing the similarities and dissimilarities between concepts in high-dimensional vector spaces: e.g., the concepts "tea" and "coffee" could be modeled with vectors that encode their similarity in that they are both drinkable substances, but also their dissimilarity in that one is made from leaves and the other from beans. To formalize the *principle of compositionality*, there have been numerous attempts to combine these LS representations into *compositional semantic* (CS) representations spanning multi-word utterances, for instance through vector averaging or multiplication (e.g., Mitchell and Lapata, 2010). However, these approaches fall short in approximating human-like compositionality, (Pavlick, 2022). Formal semantic frameworks, by contrast, fare a lot better in modeling the CS meaning of multi-word utterances. These formal semantic frameworks are typically grounded in mathematical logic, where LS meanings are modeled as functions—thereby sacrificing their conceptual richness and structure—and composition is modeled as function application (e.g., the meaning of "hot coffee" results from applying the function "hot" to the argument "coffee"). While these frameworks neatly capture CS meaning in terms of truth-conditional entailment and inference, they do not naturally capture the similarities and dissimilarities between lexical items, motivating approaches that aim to introduce distributional LS meanings into such frameworks (Garrette et al., 2014; Asher et al., 2016; Beltagy et al., 2016). While these hybrid approaches may conceptually come closest to implementing the principle of compositionality, they do often yield rather 'Frankensteinian' frameworks in which distributional and formal semantics are patched together to form a whole, while still living in distinct representational spaces, thereby not fully integrating

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

their semantic contributions.

These attempts at implementing the principle of compositionality by combining LS and CS meaning into a single semantic framework raise an important question, namely whether integrating these fundamentally different models of meaning is the right way forward. One way to address this question is to turn to how the human brain represents and constructs meaning. Advances in the neurocognition of language comprehension paint a picture supporting a perspective in which LS and CS meaning do indeed co-exist and interact, and recent neurocomputational modeling work suggests compositionality is achieved by mapping representations from an LS meaning space into a seperate space for CS meaning. Neurocognitive theory, informed by empirical and modeling results, thus suggests that LS and CS meaning do indeed inhabit distinct meaning spaces, but that they are also critically intertwined in the compositional comprehension process: incremental meaning construction involves retrieval of LS meaning, informed by the unfolding CS utterance context, which is accordingly integrated into an updated representation of the CS utterance meaning (Brouwer et al., 2012, 2017, 2021a). We therefore argue that the traditional notion of compositionality, which is grounded in syntactic combinatory rules, needs to be revised into a more dynamic notion of *compositional integration*, and we discuss the theoretical and empirical implications of this proposal.

2 The linguistic perspective: How meaning can be modeled

In the study of linguistic meaning, a variety of formal frameworks has been proposed to model meaning at the level of words, sentences, and larger discourses. While these approaches generally agree upon the principle that these levels of meaning are closely related to each other, the core phenomena studied within these frameworks vary widely, ranging from word-level similarity and conceptual structure to sentence-level entailments,

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

discourse structure and 'world knowledge'-driven inference. Attempts at implementing the principle of compositionality by integrating these approaches into a single semantic framework have proven challenging. This results in a state of affairs that suggests that LS and CS should instead be treated as complementary, but interacting, models of meaning.

2.1 Lexical semantics: Conceptual knowledge and structure

Semantic formalisms that aim to capture word-level (LS) meaning from a cognitive perspective are typically strongly grounded in the study of human semantic memory: the collection of knowledge that allows humans to not only use and understand language, but also to navigate the world around us, e.g., by recognizing and classifying objects. A core notion that these approaches aim to capture is the observation that the conceptual knowledge associated with individual words is both gradient and structured: concepts are related to each other to different degrees, which is quantified as semantic similarity (e.g., "bird" is more similar to "dog" than to "spoon"), and these relations are hierarchical in nature, in the sense that particular concepts are more general than others (e.g., "bird" subsumes both "robin" and "ostrich"). Theories of lexical meaning aim to capture this conceptual knowledge and structure by assuming semantic features as the representational currency for conceptual knowledge (McRae et al., 2005).

Semantic features constitute the dimensions of the LS representations and may take different forms (see Frisby et al., 2023). A first set of approaches intuitively conceptualizes these semantic features as identifying discrete categories or local features; for instance, the dimensions of the semantic representation of "bird" may indicate the presence/absence of features such as *has wings*, *can fly*, or *has eyes*. Each semantic representation, then, represents a vector in a high-dimensional semantic space, which can be directly compared to other representations using various vector-based metrics to quantify semantic similarity. The advantage of these approaches is that semantic similarity is not only quantifiable, but that the dimensions are also directly interpretable as independent

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

categories or features.

An alternative approach to capturing semantic features for LS is grounded in a theoretical foundation that has become known as the Distributional Hypothesis—in the formulation of J. R. Firth: "You shall know a word by the company it keeps!" (Firth, 1957, p.11). Based on the idea that "the meaning of words lies in their use" (Wittgenstein, 1953, pp. 80, 109), the Distributional Hypothesis assumes that words that occur in similar contexts will have similar meanings (see also Turney and Pantel, 2010; Clark, 2012; Erk, 2012; Lenci, 2018). This hypothesis has informed various influential implementations in which the dimensions of the resulting LS representations capture lexical co-occurrence information across linguistic contexts, i.e., sentences or documents (e.g., Latent Semantic Analysis, LSA; Landauer and Dumais, 1997, hyperspace analogue of language, HAL; Burgess, 1998, and dependency vectors, DV; Padó and Lapata, 2007). In more recent instantiations of the Distributional Hypothesis, LS vectors are word embeddings with abstract dimensions that are not directly interpretable, derived for instance from neural prediction models (e.g., `word2vec`, Mikolov et al., 2013a,b; GloVe, Pennington et al., 2014; ELMo, Peters et al., 2018; BERT, Devlin et al., 2019; GPT, Radford et al., 2019).

The resulting distributional lexical semantic (DLS) representations have been extremely successful in capturing conceptual knowledge and structure in terms of semantic similarity. This has inspired investigations into how they can be combined compositionally into utterance-level CS representations, for instance, by using vector operations as a proxy for semantic composition (Mitchell and Lapata, 2010), or by combining DLS representations into more complex structures to arrive at CS meaning (Baroni and Zamparelli, 2010; Coecke and Clark, 2011; Socher et al., 2012; Grefenstette and Sadrzadeh, 2015). While these approaches have shown some promise, for instance in modeling adjective-noun modification (Baroni et al., 2014; Vecchi et al., 2017), it has proven challenging to capture higher level semantic composition, supporting the conclusion that feature-based

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

LS representations are “good at lexical semantics, bad at composition” (Pavlick, 2022, p. 464).

2.2 Compositional semantics: The meaning of multi-word utterances

Formal semantic frameworks for CS meaning focus on modeling the construction and interpretation of phrases, sentences and multi-sentence discourses. Starting from the idea that sentences (or: propositional-level meanings) can be either true or false with respect to a state of affairs in the world, approaches in formal semantics focus on describing sentence meanings with respect to formal model structures that describe such situations. In its simplest form, a model structure is defined as a set of entities, called the *universe* U, and an interpretation function I that assigns entities from the universe (or sets thereof) to formal representations of linguistic expressions (e.g., the interpretation $I(\text{bird})$ describes the subset of entities in the universe U that are birds). Sentences can thus be assigned truth values within these model structures via a translation to some logical representation of their meaning, which in turn obtains a formal model interpretation via the interpretation function (e.g., “Tweety is a bird” is true if and only if “Tweety” refers to an entity in the universe that is also in the set of birds). Sentence meaning, then, is defined in terms of the *truth conditions* with respect to formal model structures: the constraints under which the logical representation of the sentence is assigned the truth value “true” in the model—in other words, the conditions under which the model satisfies the meaning of the sentence. Two sentences are assumed to express the same meaning if they have the same truth conditions, i.e., they are satisfied by the same models. This critically allows for a formalization of the logical entailment relation between individual sentences: Sentence A is logically entailed by sentence B if any model that satisfies the meaning of sentence B also satisfies the meaning of sentence A (e.g., the sentence “Mike paid” is logically entailed by the sentence “Mike ordered and paid”).

Approaches in semantic theory differ in terms of the logical framework that is used to

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

represent meaning as well as in terms of the complexity of the underlying model structures, which may capture, for instance, event structure (Davidson, 1969) or a notion of time (Kamp, 1980). Furthermore, traditional approaches have formalized compositional semantic construction in a static manner, assuming independent representations for lexical constituents (e.g., as lambda functions) which are then combined into compositional representations through function application (Montague, 1970). More recent semantic theorizing, however, has embraced a dynamic view toward meaning construction, emphasizing the incremental nature of linguistic processing in terms of the growth of semantic information over time (Nouwen et al., 2022).

2.2.1 Dynamic semantics: Discourse structure and composition

A dynamic semantic framework that is particularly amenable to different variations of model-theoretic complexity is Discourse Representation Theory (DRT; Kamp, 1981; Kamp and Reyle, 1993; Kamp et al., 2011). DRT is a mentalist framework for formal semantics that provides abstract representations corresponding to the types of mental representations assumed to underlie *human language comprehension*, often referred to as mental models (Johnson-Laird, 1983) or situation models (Zwaan and Radvansky, 1998). The basic meaning units in DRT are called Discourse Representation Structures (DRSs), which are formally defined as a tuple $\langle U, C\rangle$ consisting of a set of entities U and a set of conditions on these entities C. The conditions in a DRS may describe simple first-order properties or relations, but may themselves also include logical combinations of DRSs. DRSs are often visualized using box-representations such as in example (1) below, where the universe of the DRS ({x,y}) is represented in the top of the box and the conditions are described as first-order predicates over these variables:

(1) Mike called the waiter.

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

$$\begin{array}{|l|}\hline \text{x} \;\; \text{y} \\ \hline \text{x} = \text{Mike} \\ \text{waiter(y)} \\ \text{call(x,y)} \\ \hline \end{array}$$

Each DRS can be formally assigned truth conditions relative to a model structure, via either a translation to first-order logic or via an embedding function (Kamp, 1981). A critical aspect of DRT is that it formalizes meaning at the discourse rather than the sentence level; each DRS not only defines the truth conditions for a given sentence, but also provides a context for any upcoming semantic content, e.g., in terms of the referents that are available for pronominal reference. For example, a discourse in which the sentence above is continued with a novel sentence containing a referential expression, is formalized as an updated DRS in which the initial meaning representation is extended with the novel semantic information. This is effectuated as a 'merge' operation (+) between DRSs:

(2) Mike called the waiter. He did not order any food.

$$\begin{array}{|l|}\hline \text{x} \;\; \text{y} \\ \hline \text{x} = \text{Mike} \\ \text{waiter(y)} \\ \text{call(x,y)} \\ \hline \end{array} + \begin{array}{|l|}\hline \\ \hline \neg \begin{array}{|l|}\hline \text{z} \\ \hline \text{food(z)} \\ \text{order(x,z)} \\ \hline \end{array} \\ \hline \end{array} = \begin{array}{|l|}\hline \text{x} \;\; \text{y} \\ \hline \text{x} = \text{Mike} \\ \text{waiter(y)} \\ \text{call(x,y)} \\ \neg \begin{array}{|l|}\hline \text{z} \\ \hline \text{food(z)} \\ \text{order(x,z)} \\ \hline \end{array} \\ \hline \end{array} .$$

The DRS resulting from this merge operation combines the universes of both DRSs, {x,y} for the first DRS and the empty set for the second DRS, as well as their conditions.

DRT thus captures discourse-level meaning in terms of formal truth-conditional representations, while at the same time offering a dynamic semantic framework for meaning construction, in which novel semantic information is continuously merged with the

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

discourse context established so far. To arrive at these representations in a compositional manner, Muskens (1996) defines a version of DRT that employs lambda calculus to formalize how word-level meanings (formalized as functions in the form of lambda expressions) combine into sentence- and discourse-level DRS representations. Such compositional formulations, however, still assume a relatively static representation of lexical meaning, where a word like "waiter" is interpreted relative to a formal model structure as the set of entities that satisfy this predicate. This means that lexical-level similarities, as for instance modeled in distributional approaches to lexical semantics, are not naturally captured within these representations. Another important limitation of formal semantic approaches such as DRT is that that these logical frameworks do not naturally allow for capturing defeasible inferences that go beyond the literal meaning of the individual expressions—although various extensions of DRT have been proposed that do capture presuppositions and implicatures (e.g., Layered DRT; Geurts and Maier, 2013, Projective DRT; Venhuizen et al., 2018), as well as rhetorical structure (Segmented DRT; Asher and Lascarides, 2003). In particular, the interpretation of DRS representations in terms of model-derived truth conditions does not allow for capturing defeasible probabilistic inferences that reflect world knowledge-driven expectations; for instance, the inference that it is likely that "Mike" is in a "restaurant" in example (2) above. In order to capture such world knowledge-driven inferences, recent work has sought to combine insights from model-theoretic semantics with those deriving from distributional approaches to develop a framework for expectation-based semantics, which offers distributional representations of CS meaning at the level of propositions (Venhuizen et al., 2019a, 2022).

2.2.2 Expectation-based semantics: World knowledge-driven inferencing

Distributional Formal Semantics (DFS; Venhuizen et al., 2019a, 2022) is a distributional framework for meaning representation that builds on neurocognitive models of story comprehension (Golden and Rumelhart, 1993; Frank et al., 2009) to capture propositional

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

meanings in terms of co-occurrences in the world. Conceptually, DFS defines a meaning space in terms of different states-of-affairs in the world, in which propositions such as *enter(mike,bar)*, describing "Mike entering a bar", may or may not co-occur; e.g., *enter(mike,bar)* may co-occur with *order(mike,cola)*, but not with *enter(mike,restaurant)*. The DFS meaning representations that derive from this space are vectors that are compositional at the propositional level, in that meanings can be combined using logical operators, as well as probabilistic in the sense that they inherently capture the likelihood that meanings (co-)occur within the meaning space.

More formally, DFS defines meaning relative to a (finite) set of formal model structures $\mathbb{M}_{\mathbb{P}}$, which together constitute the meaning space based on a finite set of propositions $\mathbb{P}$. Each model constitutes an observation of a state of affairs in the world, in that each $M \in \mathbb{M}_{\mathbb{P}}$ is a first-order model that describes which of the propositions in $\mathbb{P}$ are true in that model. The set of models $\mathbb{M}_{\mathbb{P}}$ can thus be interpreted as a set of possible worlds, in which different constellations of propositions may co-occur (in the tradition of Carnap, 1988). The meaning of an individual proposition, then, is defined relative to this set of models (or possible worlds); that is, the meaning of a (simple or complex) proposition $p \in \mathbb{P}$ is defined by a vector $[\![p]\!]^{\mathbb{M}_{\mathbb{P}}} = \vec{v}(p)$ that assigns 1 to each $M \in \mathbb{M}_{\mathbb{P}}$ that satisfies p, and 0 otherwise (Venhuizen et al., 2022).

Critically, as propositional meaning is directly defined in terms of satisfaction with respect to formal model structures, DFS representations are fully compositional at the propositional level. This means that the meaning of any logical combination of propositions can be derived from the meaning space as operations over the underlying meaning vectors. Specifically, we can define the meaning of the negation of a given proposition p as a vector operation: $[\![\neg p]\!]^{\mathbb{M}_{\mathbb{P}}} = 1 - \vec{v}(p)$, which results in a vector that is the complement of $\vec{v}(p)$ and that assigns 0 to each $M \in \mathbb{M}_{\mathbb{P}}$ that satisfies p, and 1 otherwise. The conjunction of two propositions p and q, in turn, is defined as component-wise vector multiplication: $[\![p \wedge q]\!]^{\mathbb{M}_{\mathbb{P}}} = \vec{v}(p)\vec{v}(q)$, such that the resulting vector $\vec{v}(p \wedge q)$ assigns 1 to

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

each $M \in \mathbb{M}_{\mathbb{P}}$ that satisfies both p and q, and 0 otherwise. Together, these negation and conjunction operators allow for the derivation of any arbitrarily complex combination of propositions, as well as for definitions of existential quantification (e.g., "someone orders cola") and universal quantification ("everyone pays"); see Venhuizen et al. (2022) for details.

The set of models $\mathbb{M}_{\mathbb{P}}$ constitutes a meaning space that encodes the meaning of (complex) propositions in terms of their co-occurrence with other propositions: propositions that co-occur across a large set of models (observations of states-of-affairs in the world) will result in similar meaning vectors. Critically, while propositional meaning is defined in terms of binary vectors relative to the meaning space $\mathbb{M}_{\mathbb{P}}$, this space actually constitutes a continuous vector space $\mathbb{R}^{\mathbb{M}_{\mathbb{P}}}$. As a result, the meaning space defines meanings not only for binary propositional vectors, but also for real-valued vectors that do not directly correspond to (combinations of) propositions; rather, these vectors can be described as representing meanings that may lie in between the meanings of propositional expressions. As will become apparent below, these real-valued vectors represent sub-propositional meanings (e.g., "bartender brings" which still requires an object) that can be used to express the incremental construction of propositional-level meaning (e.g., by adding "fries" to form *bring(bartender,fries)*, which is a full proposition).

All meaning vectors that can be defined in the DFS meaning space inherently encode probabilistic knowledge about (co-)occurrence in the world that is defined by the meaning space; propositions that are true in many models can be considered to have a high probability in the world. Hence, the probability $P(a)$ of a (propositional or sub-propositional) expression a in this space is defined as follows:

$$P(a) = \frac{1}{|\mathbb{M}_{\mathbb{P}}|} \sum_{i} \vec{v}_i(a) \tag{1}$$

That is, the probability of a is defined as the fraction of models (observations) in which

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

a is satisfied. This definition can be straightforwardly extended to a definition of the conditional probability of a given b: $P(a|b) = P(a \wedge b)/P(b)$. This means that the representations in DFS allow for calculating the conditional probability of any expression in relation to all other (propositional or sub-propositional) meanings that can be defined within the meaning space. As a result, we can use this probabilistic nature of the meaning representations to quantify the extent to which expressions are inferred from each other. Specifically, if the conditional probability $P(a|b)$ equals 1 for some propositional meanings a and b, this means that a is satisfied in all the models that satisfy b; in other words, a is entailed by b ($b \vDash a$). Furthermore, by comparing the conditional probability $P(a|b)$ to the prior probability $P(a)$, the degree to which knowing b increases or decreases the certainty in a can be quantified, which gives us a notion of probabilistic inference (Venhuizen et al., 2022; Frank et al., 2009):

$$inference(a,b) = \begin{cases} \frac{P(a\,|\,b) - P(a)}{1 - P(a)} & \text{if } P(a\,|\,b) > P(a) \\ \frac{P(a\,|\,b) - P(a)}{P(a)} & \text{otherwise} \end{cases} \tag{2}$$

This inference score results in a value between -1 and 1, such that negative values indicate that a is negatively inferred from b (or: knowing b decreases the probability that a is the case) and positive values indicate that a is positively inferred from b (or: knowing b increases the probability that a is the case). Hence, an inference score of 0 indicates that a is probabilistically independent of b, an inference score of 1 indicates positive entailment ($b \vDash a$) and an inference score of -1 indicates negative entailment ($b \vDash \neg a$).

Let us turn to an example to illustrate how this mathematical machinery can be used to quantify the inferences and expectations in a concrete meaning space. Figure 1 plots the inference score for a subset of the propositions that are defined in the meaning space presented in Venhuizen et al. (2022). Propositions take the form of predicated expressions, such that *order(mike,cola)* corresponds to the meaning of "Mike orders cola". This

 © 2025 Elsevier Inc. **doi: 10.1016/bs.plm.2025.07.007**

heatmap shows the value of *inference(a,b)*, ranging from −1 (red) to +1 (green), for each propositional expression a given itself and each other propositional expression b. The green diagonal shows that each proposition is positively entailed by itself. Furthermore, certain propositions are negatively entailed by each other (e.g., *enter(mike,bar)* given *enter(mike,restaurant)*, and vice versa), which reflects the fact that in the meaning space these propositions never co-occur. All graded values reflect probabilistic inferences; for instance, *enter(mike,bar)* is inferred negatively from *order(mike,salad)*. Hence, these inferences reflect how the meaning vectors that derive from the DFS meaning space capture rich world knowledge based on propositional co-occurrences—in other words, to paraphrase the famous formulation of the Distributional Hypothesis by Firth (1957): you shall know a *proposition* by the company it keeps *in the world*.

An important observation to make here is that the inferences made within such a propositional meaning space do not directly align with word-level LS inferences informed by semantic similarity. For instance, while "bar" and "restaurant" may be elicit similar associations on the lexical level (e.g., about ordering food and drinks), the propositions in which these expressions occur are not semantically similar within the DFS meaning space, due to the (relatively) low co-occurrence of these propositions across the observations of states-of-affairs in the world. This means that the inferences that can be drawn from the DFS meaning space are distinct from those that can be drawn from lexical co-occurrences or componential analysis.

2.3 Two models of meaning?

The linguistic perspective delineates two models of meaning. On the one hand, DLS uses feature-based representations to model conceptual knowledge and structure. While these approaches do indeed successfully capture human intuitions about conceptual similarity, it has proven challenging to define compositionality over such LS representations (Pavlick, 2022). In fact, one can even raise the question if it is possible to express all of

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

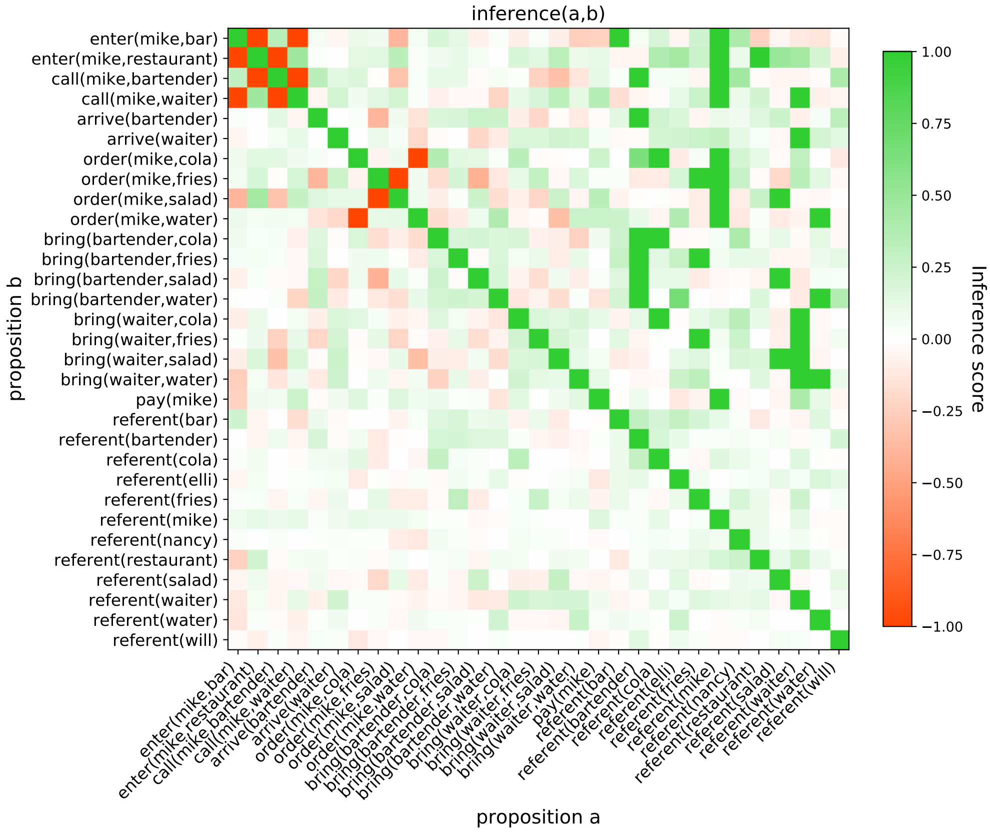

Figure 1: **Meaning space with probabilistic inferences.** Cells plot the inference score of each proposition a given each proposition b for a subset of propositions in the meaning space presented in Venhuizen et al. (2022). Bright green cells indicate positive entailment between propositions ($b \vDash a$), bright red cells indicate negative entailment ($b \vDash \neg a$), and all other intermediate cells indicate probabilistic inferences on this positive-to-negative continuum. Reproduced with permission (CC BY-NC-ND 4.0) from Venhuizen et al. (2022).

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

the complexities of compositional meaning within a meaning space for LS, of which the dimensions are assumed to represent some form of componential semantic features of individual concepts. Dynamic semantic frameworks, like DRT, on the other hand, harness formal model theory to construct CS representations that successfully capture truth-conditional entailment relations. More recent expectation-based semantic frameworks, like DFS, extend this truth-conditional approach to capturing 'world knowledge'-driven inferences in terms of probabilistic entailment relations. Neither of these formal semantic approaches to CS, however, captures the conceptual knowledge and structure that DLS approaches capture.

Various methods have been developed that aim to incorporate lexical-level distributional semantics into formal semantic frameworks (see, e.g., Coecke et al., 2010; Garrette et al., 2014; Asher et al., 2016; Beltagy et al., 2016), which for instance allow LS meaning to guide the construction of logical form for CS (Asher et al., 2016). What these approaches have in common, however, is that there remains a clear separation between the levels of representation that capture LS-derived properties (e.g., semantic similarity) and those that explain CS-derived properties (e.g., logical inference). Hence, in one way or the other, these frameworks fail to fully integrate the semantic contributions of LS and CS meaning. This raises the question if connecting these two models of meaning in a single formal semantic system is the right way forward. In what follows, we will address this question from the perspective of the neurocognition of language, and derive an architecture for incremental meaning construction that combines models of LS and CS meaning through a compositional integration process.

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

3 The neural perspective: How the brain represents meaning

The neurocognition of language comprehension is concerned with how, when, and where in the brain meaning is attributed to incoming linguistic signal as it unfolds in time. Event-Related Potentials (ERPs)—stimulus-locked, scalp-recorded voltage fluctuations caused by post-synaptic neural activity—have been instrumental in addressing questions about the how and when (see Kutas et al., 2006; Kutas and Federmeier, 2011; Hoeks and Brouwer, 2014, for reviews). ERP studies focus on systematic voltage fluctuations, referred to as *components*, which are taken to reflect specific computational operations carried out in given neuro-anatomical networks (Näätänen and Picton, 1987). Of particular salience to language comprehension are the N400 and the P600 components (see Brouwer et al., 2012; Kuperberg, 2007; Bornkessel-Schlesewsky and Schlesewsky, 2008, for reviews). Critically, the differential sensitivity of these components to aspects of LS and CS delineates a comprehension architecture in which meaning representations for LS and CS dynamically interact in the construction of compositional meaning. This dynamic interplay between LS and CS forms the core of Retrieval-Integration (RI) theory, an integrated theory of the electrophysiology of language comprehension (Brouwer et al., 2012), with an explicit cortical mapping (Brouwer and Hoeks, 2013) and neurocomputational instantiation (Brouwer et al., 2017, 2021b).

3.1 The Retrieval-Integration theory of online comprehension

RI theory, as first formulated by Brouwer et al. (2012), provides an explicit account of the processes assumed to underlie the N400 and P600 components. The N400 is a negative deflection in the ERP signal that becomes apparent 200-300ms post-word onset and peaks at about 400 ms (see Figure 2), and was first identified in response to semantically incongruous words, such as the word "socks" in "He spread the warm bread

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

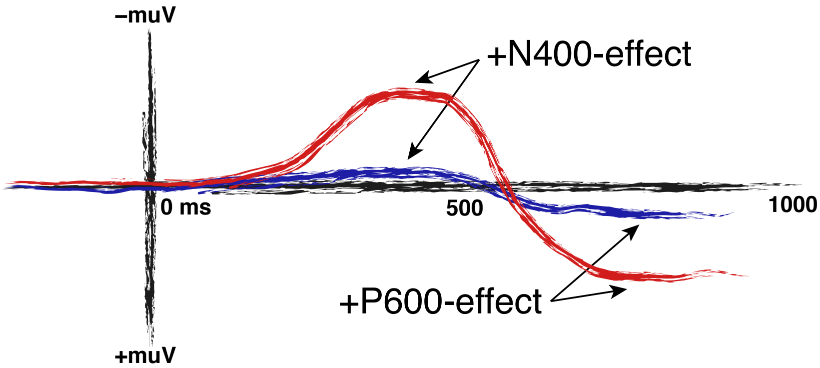

Figure 2: **N400 and P600 components in the ERP signal.** Hypothesized ERP waveform for a contrast between a target condition (red) compared to a baseline condition (blue). By convention negative voltage is plotted upwards on the y-axis. This contrast elicits both an N400 and a P600 effect for the target relative to the baseline condition, which result from the differential modulations of the N400 and P600 components in the ERP signal, respectively. Reproduced with permission (CC BY 4.0) from Brouwer and Crocker (2017).

with socks/butter" (Kutas and Hillyard, 1980). This component is, however, not just a response to an anomaly, but is in fact inversely proportional to the expectation of a word in context, such that less expected words yield larger N400 amplitudes (Kutas and Hillyard, 1984). N400 amplitude to unexpected words can, however, be attenuated if an incoming word shares semantic (Federmeier and Kutas, 1999) or orthographic features (Federmeier and Laszlo, 2009) with an expected word. Furthermore, the processes underlying the N400 are also sensitive to the semantic association of a word to its prior context (Aurnhammer et al., 2021), to the degree that strong association may override any effect of expectancy; that is, the word "socks" in the example above will not produce a larger N400 amplitude relative to "butter" when the critical sentence is embedded in a context discussing, for instance, someone trying find a fresh pair of socks before breakfast (Aurnhammer et al., 2023). Taken together, these findings pose clear constraints on the computational operations underlying the N400, leading to the now well-established perspective that the N400 is an index of the contextualized *retrieval* of feature-based LS

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

representations from long-term semantic memory, such that the more the context primes the LS features of an upcoming word, the more facilitated its retrieval and the more attenuated N400 amplitude (Kutas and Federmeier, 2000; Lau et al., 2008; Federmeier and Laszlo, 2009; van Berkum, 2009; Brouwer et al., 2012; Federmeier, 2022).

The P600, in turn, is a positive deflection in the ERP signal that starts to emerge at about 600ms post-word onset (see Figure 2), and that was first identified in response to syntactically infelicitous words, such as the word "throw" in "The spoilt child throw/throws [...]." This component is, however, not just sensitive to syntactic felicity. P600 amplitude also increases in response to structurally-induced garden-path constructions and long-distance *wh*-dependencies (Gouvea et al., 2010), semantic incongruities (Van Petten and Luka, 2012; Brouwer and Crocker, 2017), as well as a wide-range of phenomena requiring pragmatic inferencing (see Hoeks and Brouwer, 2014, for a review). Furthermore, it has recently been shown that the P600 is not just a binary reflection of well-formedness, but that its amplitude rather tracks the plausibility of a word in context in a continuous manner (Aurnhammer et al., 2023). Taken together, this is consistent with a view in which the P600 reflects the *integration* of incoming linguistic input into a CS representation of the unfolding utterance thus far, such that the more effort it takes to arrive at a coherent CS representation—in terms of construction, reorganization, and/or updating—the larger the amplitude of the P600 (Brouwer et al., 2012).

Indeed, these perspectives on the N400 as LS retrieval and the P600 as CS integration suggest that the brain harnesses two separate models of meaning for LS and CS meaning. This raises the question, however, how these meaning spaces interface in online language comprehension; that is, how do we go from the perception of words through LS to CS? RI theory offers an integrated theory of the electrophysiology of language comprehension that combines the retrieval perspective on the N400 with the integration perspective on the P600 (Brouwer et al., 2012; Brouwer and Hoeks, 2013; Brouwer et al., 2017,

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

2021b; Venhuizen and Brouwer, 2025). On RI theory, the processing of an incoming word is mechanistically conceptualized as a *process* function, that maps an acoustically or orthographically perceived *word form* in the *utterance context* in which it occurs onto a *CS representation* of utterance meaning:

$$\textit{process: (word form, utterance context)} \rightarrow \textit{CS representation} \tag{3}$$

Critically, this *process* function decomposes into a *retrieve* and *integrate* function, such that the perceived *word form* in an *utterance context* is first mapped onto a *LS representation* of word meaning:

$$\textit{retrieve: (word form, utterance context)} \rightarrow \textit{LS representation} \tag{4}$$

This contextualized retrieval of word meaning is what underlies the N400 component, and the retrieved LS representation serves as input to an *integrate* function that combines it with the *utterance context* established thus far, to produce an updated *CS representation* of utterance meaning:

$$\textit{integrate: (LS representation, utterance context)} \rightarrow \textit{CS representation} \tag{5}$$

This integration of the *LS representation* of the meaning of an incoming word with the *utterance context* underlies the P600 component. The resultant *CS representation* spanning the entire utterance will determine the *utterance context* for upcoming words; more specifically, it will serve as the *utterance context* that primes the *LS representation* associated with potential upcoming input.

RI theory thus assumes a cyclic relationship between the retrieval processes underlying the N400 and the integration processes underlying the P600. While ERPs are not directly informative about where these processes are carried out in the brain, aligning

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

insights from electrophysiology with those on the cortical organization of language—e.g., from functional Magnetic Resonance Imaging (fMRI) and lesion studies—results in a minimal functional-anatomic mapping of RI theory that further corroborates its cyclic nature (Brouwer and Hoeks, 2013). This functional-anatomic mapping is centered around the left posterior Middle Temporal Gyrus (lpMTG) as an epicenter/hub for retrieval, and the left Inferior Frontal Gyrus (lIFG) as an epicenter/hub for integration (see Figure 3a). These epicenters/hubs are connected via white matter fibers in both a dorsal pathway (dp) and a ventral (vp) pathway (see Brouwer and Hoeks, 2013, section 3.4, for further discussion). Depending on whether the input modality is spoken or written, a perceived word form enters the cortical RI cycle via either the auditory cortex (ac) or visual cortex (vc), respectively. The lpMTG then retrieves its associated LS word meaning representation, which is assumed to be stored across the association cortices, thereby generating the N400 component. The retrieved LS representation is then projected to the lIFG where it is integrated with the current utterance context to produce an updated CS utterance representation. This updated CS utterance representation in the lIFG is then connected back to the lpMTG to provide an utterance context that leads to the pre-activation/priming of (aspects of) LS representations associated with potential upcoming words (see Brouwer and Hoeks, 2013, section 4.3, for a discussion on the temporal dynamics of the communication between the lIFG and the lpMTG).

3.2 Neural meaning composition

The neurocomputational instantiation of RI theory directly implements the cortical instantiation of RI in a recurrent neural network architecture (see Figure 3b). This architecture consists of five layers, starting with an input ('ac/vc') layer at which the model receives perceived word forms. These perceived word forms are projected through a 'retrieval' (lpMTG) layer, which combines it with a top-down CS utterance context projection, from the later 'integration' (lIFG) layer, to map the perceived word form

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

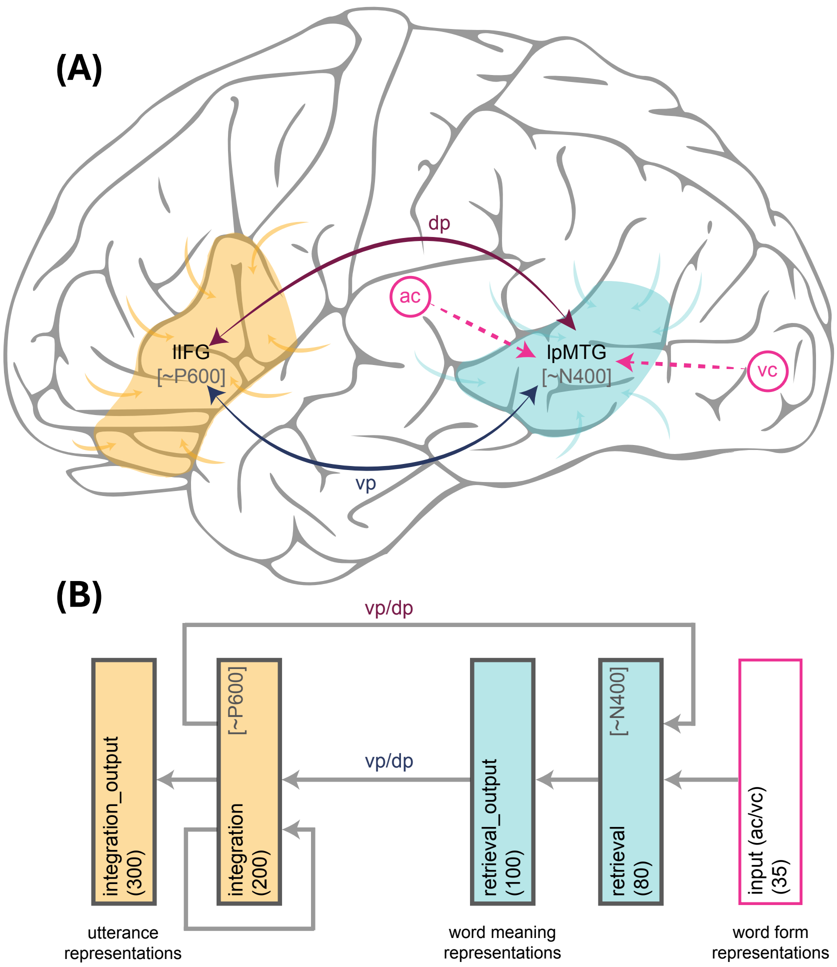

Figure 3: **Retrieval-Integration (RI) theory.** (A) Functional-anatomic instantiation of RI theory: Perceived word forms enter the RI cycle through the auditory cortex (ac) or the visual cortex (vc), depending on the input modality (spoken versus written). The left posterior Middle Temporal Gyrus (lpMTG) serves as retrieval epicenter/hub and core generator of the N400, while the left Inferior Frontal Gyrus (lIFG) is serves as integration epicenter/hub and core generator of the P600. The epicenters/hubs are connected via white matter fibers in both a dorsal pathway (dp) and ventral pathway (vp). (B) Neurocomputational instantiation of RI theory: A recurrent neural network architecture that progressively maps word forms in context onto a LS word meaning representation, and LS representations into incremental CS utterance representations. N400 amplitude is estimated as the word-induced change in activity the lpMTG layer, and P600 amplitude as the change in activity in the lIFG layer. Reproduced with permission (CC BY-NC 4.0) from Brouwer et al. (2017).

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

in context onto a LS word meaning representation in the 'retrieval_output' layer. This retrieved LS word meaning representation is then projected through a recurrent 'integration' (lIFG) layer, which combines it with the previous utterance context, to produce an updated CS utterance representation in the 'integration_output' layer. The model processes sentences on an incremental, word-by-word basis, and at each word N400 amplitude is estimated as the degree of change induced in the 'retrieval' layer, whereas P600 amplitude is estimated as the degree of change induced in the 'integration' layer. Using these explicit linking hypotheses to the N400 and P600, the model has been shown to account for key psycholinguistic processing phenomena (Brouwer et al., 2017, 2021b).

Critically, the neurocomputational instantiation of RI theory is not only explicit about its architecture and processing mechanisms, but also about the nature of the neural LS and CS representations that it assumes. The neural LS representations of word meaning are rather straightforwardly modeled as DLS representations (using the Correlated Occurrence Analogue to Lexical Semantics, COALS; Rohde et al., 2009), such that the dimensions of these vectors are proxies for componential semantic features. In the most recent instantiation of the model (Brouwer et al., 2021b), the neural CS representations are modeled using the vector representations from Distributional Formal Semantics (DFS) (Venhuizen et al., 2022). As introduced in Section 2.2.2, DFS assumes a meaning space $\mathbb{M}_{\mathbb{P}}$, consisting of set of formal model structures, such that each model $M \in \mathbb{M}_{\mathbb{P}}$ determines the truth value of each proposition $p \in \mathbb{P}$. Together these models form a continuous vector space ($\mathbb{R}^{\mathbb{M}_{\mathbb{P}}}$), and comprehension in the neurocomputational model involves navigating this vector space on a word-by-word basis to recover utterance-final propositional meaning.

This notion of comprehension as meaning-space navigation is illustrated in Figure 4. The cube in Figure 4a represents the meaning space presented in Venhuizen et al. (2022) (see also Figure 1), mapped from $|\mathbb{M}_{\mathbb{P}}| = 150$ dimensions into three dimensions (using multi-dimensional scaling, MDS). The propositional meanings that are shown represent

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

binary vectors for a subset of the propositions in $\mathbb{P}$, as well as two compositional meanings derived from combining these propositions: *enter(mike,bar)* $\wedge$ *order(mike,cola)* and *enter(mike,bar)* $\wedge$ *order(mike,fries)*. The position of these vectors relative to each other directly reflects the world knowledge in the meaning space; propositions that are likely to co-occur will be positioned closer to each other in the meaning space, and vice versa. The model learns to navigate this meaning space on a word-by-word basis, producing real-valued CS output vectors (see Figure 3b) that directly reflect world-knowledge driven inferences. Critically, the trajectory through meaning space is directly influenced by the linguistic experience that the model is exposed to, in terms of the frequency of utterance-meaning pairs encountered during training, such that the model favors trajectories for more frequently encountered word sequences (Venhuizen et al., 2019a,b).

This navigation process is illustrated in Figure 4a for the sentence prefix "Mike entered the bar, he ordered ...". After processing this sentence prefix, the model finds itself in a state that is more in line with the sentence-final meaning *enter(mike,bar)* $\wedge$ *order(mike,cola)* than with the meaning *enter(mike,bar)* $\wedge$ *order(mike,fries)*. If the sentence prefix is then continued with either "cola" or "fries", processing the word "cola" results in a more expected transition compared to processing the word "fries"—as measured by the information-theoretic notion of surprisal (Hale, 2001; Levy, 2008), which in DFS is defined as the negative logarithm of the probability of the current point in meaning space given the previous point (see Venhuizen et al., 2019a). After processing the final word, the model arrives at a point in space that approximates the intended sentence-final meaning for each sentence.

Critically, as each point in the meaning space carries its own probability in relation each other point in meaning space, the model updates its inferences about the communicated state-of-affairs on a word-by-word basis. This is illustrated in Figure 4b, which plots the inference scores (see Equation 2) for a subset of propositions pertaining to referential presuppositions, as derived from the CS representation at the output layer of

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

the model at each word of the sentence "someone called the waiter, she ordered cola". While the sentence-initial meaning vectors show no strong inferences regarding these presuppositions, the introduction of "waiter" leads to the strong inference (entailment) that a waiter is present in the described state-of-affairs. Furthermore, linguistic experience leads the model to infer the presence of female referents (*elli* and *nancy*) at the word "she". At the sentence-final word "cola", the set of probabilistic inferences reflects the 'world knowledge'-driven, non-literal interpretation that the model assigns to this sentence, namely that *elli* is a referent in the described situation (driven by the high probability of *elli* ordering *cola* in the meaning space; see Venhuizen et al., 2022 for details).

This comprehension as meaning-space navigation has several important implications. First of all, meaning composition in the model is an incremental process in which the LS meaning associated with a perceived word, in context of the CS representation established thus far, effectively triggers a transition in CS meaning-space. This transition is effectuated by the "integration" (lIFG) layer of the model, which updates its state based on its current activity pattern—its current state—and the LS of an incoming word. The degree to which this state changes as a result of processing an incoming word is an estimate of P600 amplitude in the model. Secondly, the retrieval of word meaning is effectively the activation of a word-associated LS representation in a DLS meaning-space, and this retrieval is directly affected by the state of the "integration" (lIFG) layer; that is, the "retrieval" (lpMTG) updates its state based on a word form perceived in the "ac/vc" layer, as well as the top-down state of the "integration" (lIFG) layer to retrieve the word-associated LS representation. The degree of change in this state is an estimate of N400 amplitude in the model. LS and CS meaning thus inhabit distinct meaning spaces, but are critically intertwined: compositional meaning construction involves integrating LS representations into CS space, and the current point in CS space directly affects the anticipation of aspects of upcoming LS representations.

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

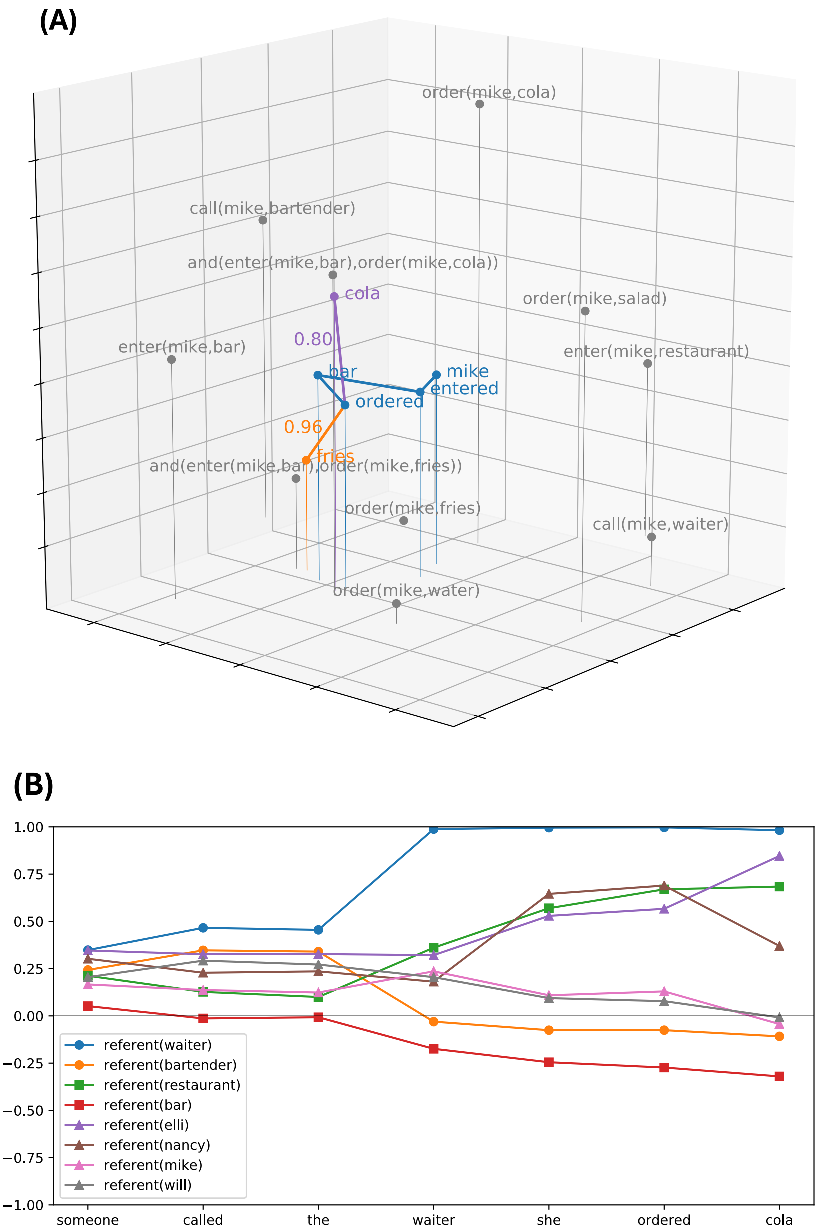

Figure 4: **Comprehension as meaning-space navigation.** (A) Three-dimensional mapping of the meaning-space presented in Venhuizen et al. (2022). The gray points show a subset of the propositions that define the meaning space, as well as two complex propositions derived from combining them. The colored points show the word-by-word trajectories for the sentences "Mike entered the bar, he ordered [cola/fries]". The numbers represented the expectancy (information-theoretic surprisal) of the sentence final words "cola" and "fries". (B) Word-by-word inference scores for propositions pertaining to referential presupposition at each word of the sentence "someone called the waiter, she ordered cola". Reproduced with permission (CC BY-NC-ND 4.0) from Venhuizen et al. (2022).

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

3.3 Decoding meaning representations from neural activity

According to RI theory, the construction of compositional utterance meaning involves a dynamic interplay between two distinct models of meaning. Conceptual meaning, on the one hand, is captured by an LS space, with representations stored across the association cortices, and the lpMTG serving has an epicenter/hub for their retrieval. Compositional utterance meaning, on the other hand, is captured by a CS space, with the lIFG serving as an epicenter/hub for the construction of an unfolding CS representation, which involves compositionally integrating LS representations into this CS space. While the neurocomputational instantiation of RI theory is both representationally explicit about LS and CS, as well as mechanistically explicit about their interplay in the compositional process, these representations and mechanisms are only simplified abstractions of those underlying comprehension in the brain. Indeed, the ultimate aim is to investigate these representations and mechanisms in the brain more directly.

Recent advances in neuroscience and artificial intelligence have led to the development of *mapping models* that do enable the direct investigation of neural meaning representation and computation in the brain through either *decoding* or *encoding* (Poldrack, 2011; King and Dehaene, 2014). These mapping models traditionally start from a set of words, LS representations for these words (of which the dimensions may or may not be directly interpretable; see Frisby et al., 2023), and neural activity patterns elicited by the perception of these words, such as individual voxel activation levels from fMRI. Decoding models then seek to accurately predict each LS dimension from these voxel activation levels, effectively yielding models that quantify the degree to which each individual voxel contributes to a particular LS dimension. Encoding models, in turn, start from the LS representations, and aim to predict each voxel activation level from the LS dimensions, yielding models that quantify the degree to which each dimension contributes to a given voxel. Critically, these encoding models can also be used for decoding, by finding the most likely cause for a pattern of observed activity, which can for instance be achieved

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

through informed search (see Tang et al., 2023, for such an approach).

While early mapping models using static LS representations—constructed using language models or human ratings—have shown that it is possible to successfully decode the meaning of words or sentences from neural activity (e.g., Mitchell et al., 2008; Pereira et al., 2018), more recent models have pushed the state-of-the-art to the decoding of continuous language by using the contextualized representations from large language models (Tang et al., 2023). Beyond practical implications of such models for brain-computer interfaces, they also provide a toolkit for directly investigating the representation and computation of meaning in the brain. However, before mapping models can be harnessed to address such fundamental questions, important methodological and and theoretical challenges need to be addressed. These challenges include the inconsistency of extant mapping results (e.g., Frisby et al., 2023) and the difficulty in reconciling these results with neurocognitive theory (e.g., compare the decoding results by Tang et al., 2023 to the cortical instantiation of RI by Brouwer and Hoeks, 2013). Furthermore, these models predominantly focus on LS and are challenged by the theoretical difficulties of the large-scale modeling of multi-word CS representations, as well as the difficulties imposed by the spatiotemporal dynamics of LS and CS representation and computation in the compositional process (see also the discussion below). While these challenges may not be straightforwardly overcome, mapping models do hold the promise to be instrumental in answering fundamental, fine-grained questions about the representation and computation of meaning in the brain.

4 The principle of compositionality revisited

The principle of compositionality assumes a close formal relationship between word-level LS and utterance-level CS meaning, since in its standard formulation, the CS meaning of an expression directly derives from the LS meanings of its constituents and the (syn-

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

tactic) rules by which they are combined (Partee, 1995). Despite this assumed close relationship, semantic theories of LS and CS meaning have developed into rather disparate fields of study. Models of LS meaning focus on representations that capture conceptual knowledge and structure, but attempts at introducing compositionality into these models—e.g., through vector averaging or multiplication (Mitchell and Lapata, 2010)—have had limited success (see Pavlick, 2022, for discussion). Models of CS meaning, on the other hand, focus on representations that capture truth-conditional entailment relations, but treat LS meaning in terms of mathematical functions, which do not capture any conceptual structure or similarity. While there have been attempts to incorporate (distributional) LS representations into such CS models, these often result in frameworks in which LS and CS representations are patched together through complex mathematical machinery, but do not fully integrate their semantic contributions (e.g., Garrette et al., 2014; Asher et al., 2016; Beltagy et al., 2016). Taken together, this raises the question of whether connecting models of LS and CS meaning in a single, unified semantic system is the right way forward.

4.1 Compositionality as a non-linear mapping between meaning spaces

Experimental findings and theoretical modeling within the neurocognition of language reveal that the human comprehension system does indeed harness both a model for LS meaning as well as a model for CS meaning. Electrophysiological research on language comprehension has shown that the N400 and the P600—the two most salient language-related components of the ERP signal—are differentially sensitive to aspects of LS and CS meaning, respectively. That is, the degree to which word-associated LS meaning is contextually anticipated has been shown to result in a reduction of N400 amplitude (e.g., Kutas, 1993; Federmeier and Kutas, 1999), while expectations regarding utterance-level CS meaning result in a reduction of P600 amplitude (e.g., Aurnhammer et al., 2023). This differential sensitivity of the N400 and P600 forms the core of the

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

Retrieval-Integration theory of language comprehension (Brouwer et al., 2012; Venhuizen and Brouwer, 2025), an integrated theory of language electrophysiology with an explicit functional-anatomic mapping (Brouwer and Hoeks, 2013) and neurocomputational instantiation (Brouwer et al., 2017, 2021b). On RI theory, the N400 component of the ERP signal indexes the retrieval of the LS meaning of a word, a process that is directly modulated by top-down CS utterance context. The P600 component, in turn, indexes the integration of this retrieved LS word meaning into an unfolding CS representation of utterance meaning. Hence, RI theory assumes LS and CS meaning to coexist and interact during language comprehension. Furthermore, the functional-anatomic mapping of RI assumes two distinct cortical epicenters/hubs, with the lpMTG serving as an epicenter/hub for the retrieval of LS representations that are assumed to be stored across the association cortices, and the lIFG as an epicenter/hub for CS meaning construction. These epicenters are wired together through dorsal and ventral white matter pathways, supporting the cyclic circuit required for top-down CS context to modulate the retrieval of incoming LS word meaning, and bottom-up LS meaning to be integrated into a representation of CS meaning.

The neurocomputational instantiation of RI theory representationally and mechanistically explicates this functional-anatomic mapping, and suggests that rather than connecting LS and CS meaning in a rule-based, formal semantic system that mathematically conflates their distinct representational currencies, compositionality may be achieved through a non-linear mapping integrating representations from an LS meaning space into a meaning space for CS; that is, the neurocomputational instantiation of RI suggests that compositionality may be an emergent epiphenomenon of the neural machinery implementing the comprehension system. Fundamentally, this is, however, still consistent with the assumption underlying the principle of compositionality that the meaning of a complex expression is determined by the meanings of the individual words that constitute the expression, and the way that they are combined.

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

Indeed, this is highly reminiscent of the way in which large language models (LLMs) construct meaning. LLMs also start from LS representations, in terms of word embeddings, which they progressively and non-linearly map into deeper, contextualized embeddings. The impressive human-like comprehension behavior of such LLMs has led to suggestions that they implement mechanisms that are highly similar to those implemented by the comprehension system in the human brain (Goldstein et al., 2022; Schrimpf et al., 2021). While such conclusions may be premature (see, e.g., Krieger et al., 2024), LLMs do offer interesting systems for further investigation. For one, the contextualized embeddings that these models construct may be the closest thing we have to wide-coverage CS representations. Hence, a better understanding of these representations by grounding them in linguistic theory and relating them to neural activity through mapping models, may further our understanding of how CS meaning is represented in the brain. Furthermore, as LLMs also start from LS representations, they serve as examples of systems that construct approximate CS representations trough non-linear mappings rather than formal, rule-based mathematical machinery, offering a means to investigate such mappings on a large scale.

4.2 Compositionality is continuous

The LS and CS models of meaning that are assumed by RI theory account for fundamentally distinct types of knowledge. The LS model is assumed to capture the conceptual structure and similarity that is associated with semantic memory. This includes conceptual knowledge regarding semantic categories and features, for instance regarding taxonomy (e.g., *is animate*, *is mammal*), function (e.g., *is edible*, *cutting tool*), and visual form (e.g., *has legs*, *made of steel*) (McRae et al., 2005). While RI theory is agnostic about the precise nature of these LS representations, the neurocomputational instantiation employs DLS representations deriving from word co-occurrences to capture conceptual similarity (based on Rohde et al., 2009; see Brouwer et al., 2017). RI theory

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

does, however, critically assume the LS meaning space to be continuous in nature; that is, since the N400 has also been shown to be sensitive in a graded manner to the degree of semantic similarity (in terms of features and/or categories; see e.g., Boddy, 1981; Bentin et al., 1985; Federmeier and Kutas, 1999), the LS meaning space should capture gradient conceptual similarity. More concretely, concepts such as *bar* and *restaurant* should have a certain degree of similarity within the LS meaning space, capturing that both have shared semantic features like *is location*, *sells food*, but also that they are associated with different features such as *has bartender* and *has waiter*, respectively.

RI theory asserts that retrieved LS meaning is integrated into an utterance-wide CS representation on a word-by-word basis. More formally, utterance representations are assumed to be dynamic in the sense that the CS meaning is captured in terms of 'context-change potential' (Nouwen et al., 2022); CS representations provide both a representation of the utterance so far, as well as a context for the retrieval of LS meaning associated with incoming words and the integration of this meaning into an updated CS representation. As such, RI assumes that the CS model allows for incremental composition of utterance-level meaning — similar to the way in which a dynamic semantic framework such as Discourse Representation Theory formalizes meaning construction.

Furthermore, the CS representations assumed by RI should not only capture literal utterance-level entailments that are the focus of standard truth-conditional semantic theories, but should also support probabilistic inferences that reflect 'world knowledge'-driven expectations; that is, since the P600 has been shown to have graded sensitivity to 'world knowledge'-driven plausibility manipulations (Aurnhammer et al., 2023), the integrative composition of CS representations should capture this gradedness. Indeed, the representations from the DFS framework (Venhuizen et al., 2022), which formalize CS meaning in the most recent computational instantiation of RI theory (Brouwer et al., 2021b), have been shown to capture graded 'world knowledge'-driven inferences as part of a high-dimensional propositional meaning space. Comprehension in the model can be

© 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

conceptualized as navigating this meaning space on a word-by-word basis, and trajectories through this space are influenced by the linguistic experience that the model is exposed to, such that gradedness can also arise from differences in utterance frequencies. In this model, CS meaning reflects propositional structure and similarity independent of feature-based LS similarity; that is, in the CS meaning space, sub-propositional meaning representations that pertain to concepts such as *bar* and *restaurant* are highly dissimilar, since the proposition *enter(mike,bar)*, for instance, leads to a probabilistic inference that *call(mike,bartender)*, while it entails the negation ¬*enter(mike,restaurant)*.

Critically, RI assumes that LS and CS meaning reside in distinct, but interacting meaning spaces, and that both of these meaning spaces are continuous in nature. As a result, the non-linear mapping from LS representations into a CS meaning space is in itself taken to be a continuous process, in that changes in contextually activated conceptual LS knowledge during comprehension will affect utterance-level CS meaning in a non-linear manner. Furthermore, the non-linear mapping from LS representations into a CS space may generalize beyond the concepts and propositional state-of-affairs that the comprehension system has experienced, thereby providing a basis for productivity and systematicity of language use, within the confines of these spaces themselves. That is, because the meaning spaces themselves are structured and capture word- and utterance-level inferences, models that describe compositional comprehension as a mapping between these spaces can map novel combinations of LS representations into the CS meaning space (productivity), and also construct novel CS meanings (systematicity), under the assumption that these meanings can be interpreted within the CS meaning space (see also Frank et al., 2009; Calvillo et al., 2021).

4.3 Compositionality is expectation-based

Expectation-based theories of language comprehension hypothesize that the comprehension system continuously generates predictions about upcoming words given the unfold-

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

ing context, be it implicitly or explicitly. On Surprisal Theory, these predictions are directly related to processing effort, such that the more unexpected an incoming word is, the higher its processing difficulty, e.g., as measured using reading times (Hale, 2001; Levy, 2008). Indeed, the cyclic nature of RI theory renders it inherently expectation-based: the top-down CS context affects both expectations about the conceptual LS meaning associated with an incoming word, as well as expectations about CS meaning resulting from integrating this LS meaning (see also Aurnhammer et al., 2021; Venhuizen and Brouwer, 2025). The degree of contextual expectations leads to graded predictions regarding N400 and P600 modulations, where the retrieval processes underlying the N400 are modulated by the degree to which LS features are pre-activated by the context, and the integration processes underlying the P600 by what can effectively be conceptualized as "comprehension-centric" surprisal—the likelihood of the current state in CS space given the previous state (Venhuizen et al., 2019a; Brouwer et al., 2021b).

The expectation-based nature of RI theory raises the question of what drives expectations about LS and CS meaning. Starting with CS meaning, expectations are directly conditioned on the current state in the CS meaning space. As each state inherently carries its own probability in the world, as well as its co-occurrence probability with other points in the meaning space, each word-induced transition in meaning space may be more or less expected within the CS space itself. In other words, world knowledge determines which states in the meaning space are positioned close to each other, thereby driving expectations regarding upcoming linguistic input. Critically, however, these transitions in meaning space are also modulated by the linguistic experience that is captured by the mapping from LS to CS representations in terms of the frequency with which certain combinations of LS meanings are mapped onto CS meanings (Venhuizen et al., 2019a). This linguistic experience reflects how often states-of-affairs are talked about in language, independent of their probability in the world. Expectations deriving from linguistic experience may often be in agreement with those deriving from

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

world knowledge, e.g., when describing a canonical situation like "John entered the cinema and ordered steak/popcorn", where the continuation "steak" is unexpected both in terms of our knowledge of the world and in terms of how frequently this situation would be described. Critically, however, world knowledge and linguistic experience may also disagree; that is, there are highly likely states-of-affairs (expected according to world knowledge) that are very uninformative and unlikely to be talked about (unexpected according to linguistic experience), e.g., "Mary drove through a green light". Indeed, it is far more likely to hear someone state that "Mary drove through a red light", as this indicates a state-of-affairs that less probable to occur in the world (assuming Mary respects traffic laws). This shows that expectations about CS meaning are thus driven by the propositional co-occurrence structure of the CS space itself, as well as by bottom-up linguistic experience (see Venhuizen et al., 2019a, for discussion).

Expectations about LS meaning, in turn, derive from an interplay between the top-down propositional co-occurrence structure of the CS space, bottom-up linguistic experience, as well as world knowledge-driven conceptual structure of semantic memory. First of all, the mapping of word form onto a LS meaning representation—i.e., retrieval of word meaning—is modulated by top-down CS context, meaning that similar CS contexts will lead to the anticipation of similar LS meanings. Which LS meanings are anticipated in a given CS context, however, is determined by linguistic experience; that is, it is linguistic experience that shapes the relative strength of the association between a given CS context and specific LS meanings. Finally, LS meanings that are positioned relatively close in the conceptual meaning space will share activation patterns and may therefore also influence lexical-level expectations. Hence, expectations about both LS and CS meaning are modulated by the linguistic experience that the system is exposed to, as well as both conceptual and propositional world knowledge (see also Troyer and Kutas, 2020a,b, for direct empirical investigations of the influence of world knowledge on word processing).

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

4.4 Compositionality is spatiotemporally extended

The functional-anatomic mapping of RI theory assumes a spatial segregation between the epicenters/hubs for retrieval and integration in terms of the lpMTG (plus association cortices) and lIFG, respectively (Brouwer and Hoeks, 2013). This spatial segregation can be addressed using mapping models, as discussed in Section 3.3. At a bare minimum, this means that mapping model investigations into LS meaning, CS meaning, and the compositional process should honor this segregation: the lpMTG and association cortices are predicted to be more involved in LS retrieval, whereas the lIFG is predicted to be more focally involved in CS integration. This state of affairs is, however, further complicated by the temporal dynamics of the assumed retrieval and integration processes; that is, the retrieval and integration processes are known to be active simultaneously, leading the N400 and P600 to spatiotemporally overlap in the scalp-recorded ERP signal (see Delogu et al., 2019, 2021; Brouwer et al., 2021a; Delogu et al., 2025). Beyond complications for interpreting this scalp-recorded ERP signal (see Brouwer and Crocker, 2017, for discussion), this implies that the compositional process is also spatiotemporally extended. As a consequence, mapping models should take both the spatial and temporal dynamics of the compositional process into account. Going forward, we should thus disentangle LS and CS representation in space, by building mapping models that target data from neuroimaging methods with high spatial resolution such as fMRI, as well as in time, through mapping models targeting data from neuroimaging methods with high temporal resolution such as electroencephalography (EEG). To synthesize the results on space and time, mapping models could be complemented by neurocomputational models that explicate the spatiotemporal dynamics underlying compositionality in comprehension, such as a temporally-extended version of the neurocomputational instantiation of RI theory (see Brouwer et al., 2017, section 5.4, for discussion).

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

5 Conclusions

Formal modeling approaches in linguistic theory and the neurocognition of language comprehension are both concerned with the question of how meaning is represented and constructed from linguistic signal. The principle of compositionality, which assumes that the meaning of a complex expression is defined as a function of the meaning of its parts and the way they are combined, has long been a hallmark of formal semantic approaches. Extant models of semantic theory, however, focus on either capturing lexical semantic meaning in terms of the conceptual knowledge and structure, or compositional meaning in terms of truth-conditional entailments and inferences. Attempts at directly integrating these models of lexical semantics with models of utterance-level compositional semantics—to formalize a single semantic framework for compositional meaning representation and construction—have proven challenging, and question the validity of this endeavor. On the other hand, recent neurocognitive theorizing and modeling reveals an architecture for language comprehension that assumes Retrieval-Integration cycles, in which word-by-word processing involves the retrieval of lexical semantic word meaning from long-term memory, and the integration of these lexical semantic meanings into a coherent representation of compositional semantic utterance meaning.

Combining insights from linguistic theory regarding the nature of the representations for lexical semantics and utterance-level compositional semantics with the computational mechanisms assumed to underlie Retrieval-Integration cycles, paints a picture in which compositional meaning construction harnesses two separate, but interacting models of meaning—one for lexical semantics and one for compositional semantics—that dynamically interact during the incremental process of word-by-word meaning construction. Within this architecture, compositionality arises from a non-linear mapping of lexical semantic representations into a space for utterance-level compositional meaning. This results in a notion of *compositional integration*, which emphasizes the continuous nature

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

of the compositional process and its underlying representations, the expectation-based dynamics of word-by-word meaning composition, as well as the observation that incremental meaning construction is a spatiotemporally-extended process in the brain. This novel perspective on compositionality—centered around two models of meaning—thus combines insights from linguistic and neurocognitive theory, and serves as a starting point for more integrative, interdisciplinary approaches towards modeling the representation and computation of the meaning of words, sentences, and larger discourses.

References

Asher, N. and Lascarides, A. (2003). *Logics of Conversation.* Cambridge University Press.

Asher, N., Van de Cruys, T., Bride, A., and Abrusán, M. (2016). Integrating type theory and distributional semantics: a case study on adjective–noun compositions. *Computational Linguistics*, 42(4):703–725.

Aurnhammer, C., Delogu, F., Brouwer, H., and Crocker, M. W. (2023). The P600 as a continuous index of integration effort. *Psychophysiology*, 60:e14302.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., and Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, 16(9):e0257430.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014). Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology (LiLT)*, 9:241–346.

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010*

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

Conference on Empirical Methods in Natural Language Processing, pages 1183–1193. Association for Computational Linguistics.

Beltagy, I., Roller, S., Cheng, P., Erk, K., and Mooney, R. J. (2016). Representing meaning with a combination of logical and distributional models. *Computational Linguistics*, 42(4):763–808.

Bentin, S., McCarthy, G., and Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and clinical Neurophysiology*, 60(4):343–355.

Boddy, J. (1981). Evoked potentials and the dynamics of language processing. *Biological Psychology*, 13:125–140.

Bornkessel-Schlesewsky, I. and Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain research reviews*, 59(1):55–73.

Brouwer, H. and Crocker, M. W. (2017). On the proper treatment of the N400 and P600 in language comprehension. *Frontiers in Psychology*, 8:1327.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., and Hoeks, J. C. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, 41:1318–1352.

Brouwer, H., Delogu, F., and Crocker, M. W. (2021a). Splitting event-related potentials: Modeling latent components using regression-based waveform estimation. *European Journal of Neuroscience*, 53(4):974–995.

Brouwer, H., Delogu, F., Venhuizen, N. J., and Crocker, M. W. (2021b). Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12:615538.

 © 2025 Elsevier Inc. **doi: 10.1016/bs.plm.2025.07.007**

Brouwer, H., Fitz, H., and Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, 1446:127–143.

Brouwer, H. and Hoeks, J. C. (2013). A time and place for language comprehension: Mapping the N400 and the P600 to a minimal cortical network. *Frontiers in Human Neuroscience*, 7:758.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, 30(2):188–198.

Calvillo, J., Brouwer, H., and Crocker, M. W. (2021). Semantic systematicity in connectionist language production. *Information*, 12(8):329.

Carnap, R. (1988). *Meaning and necessity: A study in semantics and modal logic*, volume 30. University of Chicago Press.

Chomsky, N. (1957). *Syntactic Structures.* Mouton & Co., N.V., 's-Gravenhage, Netherlands.

Clark, S. (2012). Vector space models of lexical meaning. In Lappin, S. and Fox, C., editors, *Handbook of Contemporary Semantics–second edition*, pages 493–522. Wiley-Blackwell.

Coecke, B., Sadrzadeh, M., and Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

Coecke, M. S. B. and Clark, S. (2011). Mathematical foundations for a compositional distributional model of meaning. In *Festschrift for Joachim Lambek*, volume 36 of *Linguistic Analysis*, pages 345–384. Linguistic Analysis.

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

Davidson, D. (1969). The individuation of events. In *Essays in honor of Carl G. Hempel: A tribute on the occasion of his sixty-fifth birthday*, pages 216–234. Springer.

Delogu, F., Aurnhammer, C., Brouwer, H., and Crocker, M. W. (2025). On the biphasic nature of the n400-p600 complex underlying language comprehension. *Brain and Cognition*, 186:106293.

Delogu, F., Brouwer, H., and Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, 135:103569.

Delogu, F., Brouwer, H., and Crocker, M. W. (2021). When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension. *Brain Research*, 1766:147514.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Federmeier, K. D. (2022). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology*, 59(1):e13940.

Federmeier, K. D. and Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, 41(4):469–495.

Federmeier, K. D. and Laszlo, S. (2009). Time for meaning: Electrophysiology provides

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

insights into the dynamics of representation and processing in semantic memory. *Psychology of learning and motivation*, 51:1–44.

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

Frank, S. L., Haselager, W. F., and van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition*, 110(3):358–379.

Frisby, S. L., Halai, A. D., Cox, C. R., Ralph, M. A. L., and Rogers, T. T. (2023). Decoding semantic representations in mind and brain. *Trends in Cognitive Sciences*, 27(3):258–281.

Garrette, D., Erk, K., and Mooney, R. (2014). A formal approach to linking logical form and vector-space lexical semantics. In *Computing meaning*, pages 27–48. Springer.

Geurts, B. and Maier, E. (2013). Layered Discourse Representation Theory. In Capone, A., Piparo, F. L., and Carapezza, M., editors, *Perspectives on Linguistic Pragmatics*, pages 311–327. Springer International Publishing.

Golden, R. M. and Rumelhart, D. E. (1993). A parallel distributed processing model of story comprehension and recall. *Discourse processes*, 16(3):203–237.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Gouvea, A. C., Phillips, C., Kazanina, N., and Poeppel, D. (2010). The linguistic processes underlying the p600. *Language and cognitive processes*, 25(2):149–188.

Grefenstette, E. and Sadrzadeh, M. (2015). Concrete models and empirical evaluations

 © 2025 Elsevier Inc. **doi: 10.1016/bs.plm.2025.07.007**

for the categorical compositional distributional model of meaning. *Computational Linguistics*, 41(1):71–118.

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8, Stroudsburg, PA. Association for Computational Linguistics.

Hoeks, J. C. J. and Brouwer, H. (2014). Electrophysiological research on conversation and discourse processing. In Holtgraves, T. M., editor, *The Oxford Handbook of Language and Social Psychology*, pages 365–386. New York: Oxford University Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Harvard University Press, Cambridge, MA.

Kamp, H. (1980). Some remarks on the logic of change, Part I. In Rohrer, C., editor, *Time, Tense, and Quantifiers: Proceedings of the Stuttgart Conference on the Logic of Tense and Quantification,*, pages 135–180. Max Niemeyer Verlag, Berlin, New York.

Kamp, H. (1981). A theory of truth and semantic representation. In Groenendijk, J. A. G., Janssen, T. M. V., and Stokhof, M. B. J., editors, *Formal Methods in the Study of Language, Proceedings of the Third Amsterdam Colloquium*, pages 277–322, Amsterdam. Mathematisch Centrum.

Kamp, H. and Reyle, U. (1993). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and Discourse Representation Theory.* Kluwer, Dordrecht.

Kamp, H., van Genabith, J., and Reyle, U. (2011). Discourse Representation Theory. In Gabbay, D. M. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 15, pages 125–394. Springer Netherlands.

 © 2025 Elsevier Inc. **doi: 10.1016/bs.plm.2025.07.007**

King, J.-R. and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, 18(4):203–210.

Krieger, B., Brouwer, H., Aurnhammer, C., and Crocker, M. W. (2024). On the limits of llm surprisal as functional explanation of erps. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146:23–49.

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and cognitive processes*, 8(4):533–572.

Kutas, M. and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4(12):463–470.

Kutas, M. and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62:621–647.

Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Kutas, M., van Petten, C., and Kluender, R. (2006). Psycholinguistics electrified II: 1994–2005. In Traxler, M. J. and Gernsbacher, M. A., editors, *Handbook of Psycholinguistics, 2nd Edition*, pages 659–724. Elsevier, New York.

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

Landauer, T. K. and 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.

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, 9(12):920–933.

Lenci, A. (2018). Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

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

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

Montague, R. (1970). Universal grammar. *Theoria*, 36(3):373–398.

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

Muskens, R. (1996). Combining Montague semantics and discourse representation. *Linguistics and Philosophy*, 19(2):143–186.

Näätänen, R. and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24(4):375–425.

Nouwen, R., Brasoveanu, A., van Eijck, J., and Visser, A. (2022). Dynamic Semantics. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational linguistics*, 33(2):161–199.

Partee, B. H. (1995). Lexical semantics and compositionality. In Gleitman, L., Liberman, M., and Osherson, D. N., editors, *An Invitation to Cognitive Science, Volume 1: Language*, pages 311–360. The MIT press, Cambridge, MA, 2nd edition.

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, 8:447–471.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., and Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M., Ji, H., and

 © 2025 Elsevier Inc. **doi: 10.1016/bs.plm.2025.07.007**

Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, 72(5):692–697.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Rohde, D. L. T., Gonnerman, L. M., and Plaut, D. C. (2009). An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science*, pages 1–33.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211. Association for Computational Linguistics.

Tang, J., LeBel, A., Jain, S., and Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866.

Troyer, M. and Kutas, M. (2020a). Harry potter and the chamber of what?: The impact of what individuals know on word processing during reading. *Language, cognition and neuroscience*, 35(5):641–657.

 © 2025 Elsevier Inc. **doi: 10.1016/bs.plm.2025.07.007**

Troyer, M. and Kutas, M. (2020b). To catch a snitch: Brain potentials reveal variability in the functional organization of (fictional) world knowledge during reading. *Journal of Memory and Language*, 113:104111.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

van Berkum, J. J. A. (2009). The 'neuropragmatics' of simple utterance comprehension: An ERP review. In Sauerland, U. and Yatsushiro, K., editors, *Semantics and Pragmatics: From experiment to theory*, pages 276–316. Palgrave Macmillan, Basingstoke.

Van Petten, C. and Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and erp components. *International journal of psychophysiology*, 83(2):176–190.

Vecchi, E. M., Marelli, M., Zamparelli, R., and Baroni, M. (2017). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive science*, 41(1):102–136.

Venhuizen, N. J., Bos, J., Hendriks, P., and Brouwer, H. (2018). Discourse semantics with information structure. *Journal of Semantics*, 35(1):127–169.

Venhuizen, N. J. and Brouwer, H. (2025). Referential retrieval and integration in language comprehension: An electrophysiological perspective. *Psychological Review*.

Venhuizen, N. J., Crocker, M. W., and Brouwer, H. (2019a). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, 56(3):229–255.

Venhuizen, N. J., Crocker, M. W., and Brouwer, H. (2019b). Semantic entropy in language comprehension. *Entropy*, 21(12):1159.

 © 2025 Elsevier Inc. doi: 10.1016/bs.plm.2025.07.007

Venhuizen, N. J., Hendriks, P., Crocker, M. W., and Brouwer, H. (2022). Distributional formal semantics. *Information and Computation*, 287:104763. Special Issue: Selected Papers from WoLLIC 2019, the 26th Workshop on Logic, Language, Information and Computation.

Wittgenstein, L. (1953). *Philosophical Investigations.* Basil Blackwell, Oxford.

Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185.

 © 2025 Elsevier Inc. **doi: 10.1016/bs.plm.2025.07.007**