frontiers | Frontiers in Language Sciences

TYPE Review
PUBLISHED 10 March 2023
DOI 10.3389/flang.2023.1096110

OPEN ACCESS

EDITED BY
William Matchin,
University of South Carolina, United States

REVIEWED BY
Ellen F. Lau,
University of Maryland, College Park, United States
Elliot Murphy,
University of Texas Health Science Center at Houston, United States

*CORRESPONDENCE
Lia Călinescu
✉ lia.calinescu@ntnu.no

SPECIALTY SECTION
This article was submitted to Neurobiology of Language, a section of the journal Frontiers in Language Sciences

RECEIVED 11 November 2022
ACCEPTED 21 February 2023
PUBLISHED 10 March 2023

CITATION
Călinescu L, Ramchand G and Baggio G (2023) How (not) to look for meaning composition in the brain: A reassessment of current experimental paradigms.
Front. Lang. Sci. 2:1096110.
doi: 10.3389/flang.2023.1096110

How (not) to look for meaning composition in the brain: A reassessment of current experimental paradigms

Lia Călinescu[1]*, Gillian Ramchand[2] and Giosuè Baggio[1]

[1]Language Acquisition and Language Processing Lab, Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway, [2]Department of Language and Culture, Center for the Advanced Study of Theoretical Linguistics, The Arctic University of Norway, Tromsø, Norway

When we use language, we draw on a finite stock of lexical and functional meanings and grammatical structures to assign meanings to expressions of arbitrary complexity. According to the Principle of Compositionality, the meanings of complex expressions are a function of constituent meanings and syntax, and are generated by the recursive application of one or more *composition operations*. Given their central role in explanatory accounts of human language, it is surprising that relatively little is known about how the brain implements these composition operations in real time. In recent years, neurolinguistics has seen a surge of experiments investigating when and where in the brain meanings are composed. To date, however, neural correlates of composition have not been firmly established. In this article, we focus on studies that set out to find the correlates of linguistic composition. We critically examine the paradigms they employed, laying out the rationale behind each, their strengths and weaknesses. We argue that the still blurry picture of composition in the brain may be partly due to limitations of current experimental designs. We suggest that novel and improved paradigms are needed, and we discuss possible next steps in this direction. At the same time, rethinking the linguistic notion of composition, as based on a tight correspondence between syntax and semantics, might be in order.

KEYWORDS

composition, compositionality, semantics, experimental paradigms, brain, methodology

1. Introduction

Linguistic communication rests on our capacity to combine the meanings of morphemes and words into complex semantic structures. This basic property of language has been a central concern in linguistics for decades. More recently, it has attracted the attention of neurolinguists, as the need to understand its neurobiological underpinnings has become pressing. Research on "composition," "unification," "combinatorics," or "integration" is now common in cognitive neuroscience. Yet, the mechanisms by which meaning is composed in the brain remain at present elusive: neural correlates of composition, invariant across experiments using different paradigms and methods, have not yet been established. The delay in our understanding of composition in the brain may partly stem from limitations inherent in the paradigms used so far: we will argue that *none of them currently affords the direct comparisons between conditions that could reveal a correlate or signature of composition*. As we review these paradigms, we will identify a number of requirements that future experiments should meet to achieve that goal.

But how should composition be defined? At the computational level (Marr, 1982; Baggio, 2018) in formal semantics and adjacent fields, composition is the operation that, for any given complex expression *E*, takes as input *E*'s immediate constituent meanings and *E*'s constituent structure and outputs *E*'s meaning (Heim and Kratzer, 1998). Compositionality is the idea that there is a strong parallelism, or one-to-one correspondence, between the operations that build syntactic structures and meaning composition: each application of meaning composition mirrors the application of syntactic structure-building operations. In the Minimalist Program, Merge is used to derive hierarchical constituent structure by recursively forming sets of syntactic objects in pairs (Adger, 2003). In standard versions of formal semantics, composition amounts to the "saturation" of "unsaturated" meanings (e.g., a verb by its arguments) *via* the operation known as Functional Application, where a function is applied to arguments of appropriate type (Heim and Kratzer, 1998). All these operations are characterized *atemporally* in any formal system that strives to model the syntax and semantics of a language. At Marr's (1982) algorithmic and implementational levels of analysis, instead, these operations are modeled as *processes* unfolding in time. Our focus is on studies investigating *local* composition of *linearly adjacent* functional or lexical items, where there is a direct correspondence between logic and time, or between the deployment of composition at the computational level and its algorithmic and neural execution. This correspondence becomes more complex with non-adjacent constituents, which pose specific problems for theories and experiments. Moreover, our focus will be on on-line language *comprehension*, not on production: little is known (and perhaps can be known experimentally) about whether and how meanings are *composed* during early stages of conceptualization and message generation.

The question of the neural bases of composition and Compositionality has only recently been brought to the foreground of research. But this move did not provide the hoped-for advancements: the way meaning is composed in the brain remains an unsolved problem (Pylkkänen, 2019). Current experiments have not been based on paradigms that reliably vary the presence vs. absence of composition. At a minimum, the field has not benefited from enough discussion on whether accepted and presently used paradigms achieve the intended aims. This paper tries to fill this gap. We will not focus so much on the *results* of each study: those cannot be confidently interpreted unless the validity of paradigms is thoroughly assessed. Published research may report spatio-temporal activity that differs between conditions, but those effects may not entirely reflect the processes of interest, *if the baseline conditions cannot fully prevent composition*. Furthermore, it is not obvious that limitations of current paradigms can be mitigated by using higher-resolution neural recordings or more advanced methods for analyzing data. Progress is needed on several fronts simultaneously: here we concentrate on the paradigms front.

We should emphasize that the same paradigm or design may be inadequate for studying composition and perfectly suitable for other aims, for example the identification of brain signatures of syntactic or semantic processing complexity. On the one hand, this implies that some paradigms are "almost good enough", in that they successfully target processes closely associated with composition. On the other hand, this should remind readers that our aim is not to disqualify certain paradigms, designs, studies, or research programs *as such*, or even as viable approaches to the experimental study of syntax and semantics in the brain, but only and specifically as they relate to syntax-driven meaning composition, as defined above. We will thus discuss studies that manipulate the inputs of composition (constituents meanings and syntax) and ask whether the chosen conditions are adequate to identify, upon subtraction or comparison of neural responses, a correlate or signature of meaning composition. Even if a paradigm was not originally or primarily intended to study composition, we can still ask whether it *can be leveraged* to do that.

Research on composition makes the rather plausible assumption that, for meanings that *are* regimented by Compositionality, we should be able to identify experimentally neural events that instantiate composition in comparison to conditions where the requirements of composition are *not* met, because constituent structure or meaning cannot be derived or the meanings of the parts are unavailable. The challenge is indeed to utilize control or baseline conditions that can prevent the system from engaging in composition. Syntactic and semantic processing are however correlated. One problem for isolating composition in the brain is that studies that vary the structure of a stimulus tend to vary its meaning as well (Pylkkänen, 2019), and covarying neural signals can be difficult to disentangle. A second challenge is that composition is correlated or co-occurring with *other* processes, including non-strictly-compositional processes, like conceptual combination, pragmatic processing, inference etc. (Baggio et al., 2016). Thirdly, linguistic theories and processing models do not yet fully agree on the steps by which structure and meaning are built, and linking hypotheses that can effectively connect levels of analysis and guide experimental research are scarce (Baggio et al., 2012a; Pylkkänen and Brennan, 2019; Baggio, 2020).

Most paradigms that have been used to study compositional processes vary the presence or absence of syntax or lexical semantics. By subtracting the compositional and baseline responses, they attempt to isolate only that which differs between the two: neural events associated with syntax-driven meaning composition. We discuss paradigms that use this approach (Section 2) or that exploit particularities of languages to vary semantics while keeping structure constant or vice versa (Section 3). Although we take structure to be an essential ingredient in meaning composition, studies that attempt to isolate composition should look not just at structure building *per se*, but at the derivation of *meaning* guided by structure. Thus, we will also consider studies investigating syntactic composition that used stimuli with compositional meaning (e.g., Pallier et al., 2011). Instead, experiments on syntactic structure in designs where meaning is absent, such as artificial grammars, will not be considered, along with studies using classical semantic or syntactic violations (e.g., Ni et al., 2000; Friederici et al., 2004). These designs are not suitable for isolating syntactic or semantic composition. They include well-formed sentences where semantic or syntactic constraints are violated on single words, but no comparisons that can reveal syntactic or semantic composition. Furthermore, the brain might still attempt to derive meaning in anomalous sentences, even though that meaning may not be licensed by

the structure of the input or may conflict with conceptual or world knowledge or pragmatic constraints (Pylkkänen et al., 2011). Violations may also trigger repair mechanisms, after which composition could theoretically still apply.

2. The beaten path: Three classical paradigms

2.1. Scrambling linear order

Sentences are not linear sequences of words, but recursive, hierarchical combinations of words and phrases. One widely used paradigm compares syntactically well-formed and meaningful expressions with stimuli where the linear order of words is *scrambled*. This manipulation is assumed to prevent the formation of syntactic structures at all levels of the hierarchy (phrases, clauses), thus disrupting compositional operations. Experiments using this approach can be separated into two groups, based on the "size" of the linguistic structures used for comparison: sentences or narratives.

2.1.1. Well-formed sentences vs. lists of words

One type of paradigm compares well-formed sentences with word lists where the linear order is broken, and thus syntactic hierarchies and complex meanings cannot be formed (Hashimoto and Sakai, 2002; Kuperberg et al., 2000). This paradigm would seem to target Compositionality directly: the meaning of a sentence is not just given by the meanings of constituents; syntactic structure plays a role, too. One assumption behind this paradigm is that lists and sentences only differ in one respect: syntactic structure. As we will see in this section, however, that assumption does not always hold.

This paradigm has been used in combination with other baselines, such as pseudoword sentences ("Jabberwocky") and pseudoword lists, to be discussed below. Word lists and scrambled sentences are used in fMRI studies to identify broad patterns of activation for language processing (Fedorenko et al., 2010) and isolate specific functional components of language (e.g., syntactic and semantic processing). In these studies, fully well-formed meaningful sentences are compared either to scrambled versions of the same sentences, where the same content and function words are presented in random order, or to lists of words not present in the original sentences. The assumption is that processes engaged at the single word level (e.g., lexical retrieval) are equally present in lists and sentences, so subtraction (sentence–list) will isolate neural responses that differ between conditions, such as a putative neural correlate of syntax-driven composition. Across studies, there is variability in how baseline conditions with word lists are built: as we will see soon, this is indirect proof of the challenges that arise when constructing stimuli in this paradigm.

Sentences compared to unstructured lists involve the construction of sentence structure *and* meaning. Some studies have thus included manipulations of meaning to tease them apart. Vandenberghe et al. (2002) used PET with a blocked design comparing sentences to lists to determine the contribution of syntax to composition. The lists used scrambled content and function words from the sentences. Similarly, Humphries et al. (2006) ran an fMRI study using semantically congruent sentences (1a) and lists (1b):

(1) a. The man on a vacation lost a bag and wallet
b. On vacation lost then a and bag wallet man then a

Both studies used semantic manipulations to disentangle compositional semantics from syntax. Semantically "random" sentences (1c) were compared with semantically "random" lists (1d); from Humphries et al. (2006):

(1) c. The freeway on a pie watched a house and a window
d. A ball the a the spilled librarian in sign through fire

The incongruent condition (1c) is intermediate between a congruent sentence and a list: it has structure, but meaning is deviant. Incongruent sentences should control syntactic structure and lexical retrieval, but differences related to contextual activation of specific lexical items still exist between these conditions. Further, plausibility or meaningfulness manipulations may not prevent composition. Semantically anomalous sentences used in these studies appear felicitous up to the first few words, allowing participants to initially compose meaning ("Youths resented a sketch of the forest"). The results indicate that a subset of regions active for sentences is also active for anomalous sentences and lists, as if the brain engaged in composition in all conditions, albeit possibly to different extents or at different positions across items.

Goucha and Friederici (2015) compared well-formed and meaningful sentences (2a) with well-formed incongruent sentences (2b) and scrambled lists of unrelated words (2c):

(2) a. The complexity of the regulations had shocked the unhappy kingdom
b. The vicinity of the constipation had ironed the uncanny wisdom
c. Vicinity the of had constipation wisdom ironed uncanny the

In Goucha and Friederici (2015), lexical information is matched across sentences and lists. To reduce the risk of incidental syntactic structure building in word lists, Humphries et al. (2006) created lists by randomly sampling function words from the stimulus set and by replacing them in sentences before shuffling their word order. Lexical content cannot be matched exactly, but randomly picked function words might be less likely to combine with the given content words. Even so, this is unlikely to fully block syntactic processing.

Another scrambling approach was used by Kaufeld et al. (2020):

(3) a. [Bange helden] [plukken bloemen] en de [bruine vogels] [halen takken]
[*Timid heroes*] [*pluck flowers*] *and the* [*brown birds*] [*gather branches*]
b. [helden bloemen] [vogels takken] de en [plukken halen] [bange bruine]
[*heroes flowers*] [*birds branches*] *the and* [*pluck gather*] [*timid brown*]

They found increased neural tracking, at the phrase frequency, for sentences (3a) vs. lists (3b), which suggests the brain is building hierarchical structure for sentences but not for lists. This could be taken to indicate that meaning composition too is happening only for sentences, tracking hierarchical structure. But stringing together locally words from the same category in lists could lead to compounding attempts (N-N), or could engage other syntactically viable modes of combination (e.g., adjective stacking, "bange bruine"). This issue is not specific to this study, but applies widely to lists paradigms. Composition may then occur in both sentences and lists, at least locally.

Another variant of this paradigm disrupts syntactic structure *parametrically*, resulting in conditions with different degrees of scrambling. Pallier et al. (2011) studied the neural mechanisms of hierarchical structure building using stimuli with five levels of scrambling. These varied in the size of the constituents, ranging from a full sentence (4a) to a word list (4f), with lists of constituents of different sizes in between, 6 to 2 words, (4b) to (4e):

(4) a. I believe that you should accept the proposal of your new associate
b. [the mouse that eats our cheese] [two clients examine this nice couch]
c. [mayor of the city] [he hates this color] [they read their names]
d. [solving a problem] [repair the ceiling] [he keeps reading] [will buy some]
e. [looking ahead] [important task] [who dies] [his dog] [few holes] [they write]
f. thing very tree where of watching copy tested they states heart plus

As constituents were extracted from the sentence condition and concatenated randomly, lexical material was matched across conditions, but not within each items set. Activation was modulated by constituent size in the left superior temporal sulcus (STS) and inferior frontal gyrus (LIFG partes triangularis and orbitalis). In a replication study, Shain et al. (2021) suggest that these effects may not reflect syntactic structure building, but the fact that shorter constituents may not fully engage the language network. Larger chunks may then be easier to recognize by the language network as stimuli to be processed.

Parametric variation of constituent size can be a way of overcoming the poorer temporal resolution of BOLD fMRI and can be used to track how composition unfolds step by step as structure and meaning are built. However, as noted by Grodzinsky et al. (2021), these designs are not without issues. The conditions do not form minimal pairs: e.g., there are additional differences in category labels and number of structural units between them. A similar study is Matchin et al. (2017), who aimed to dissociate the effects of bottom-up syntactic computations from those of top-down predictions, by comparing lists of words ("rabbit the could extract protect") to lists of two-word phrases ("the fencer the baby the bill") and full sentences ("the poet will recite a verse"). Zaccarella et al. (2017) matched as much as possible semantic content between conditions, comparing sentences ("The ship sinks") to prepositional phrases that contained a matched noun ("on the ship"). The word list baseline contained a further control measure: the nouns in the lists were in the same positions as in the sentence or phrase ("stem *ship* juice"; "leek mouth *ship*"). Matchin et al. (2017) and Zaccarella et al. (2017) find effects in the left IFG and posterior STS (pSTS) for syntactic structure building, but only the former study reports effects in left pSTS for sentences and phrases.

Mollica et al. (2020) compare in an fMRI study well-formed sentences (5a) to scrambled sentences with 1 swap (5b), 3 (5c), 5 (5d) or 7 swaps (5e), and a list of content words:

(5) a. on their last day they were overwhelmed by farewell messages and gifts
b. on their last day they were overwhelmed by farewell *and messages* gifts
c. on their last *they day* were overwhelmed *farewell by* and *messages gifts*
d. on their last day *were overwhelmed they farewell messages by gifts and*
e. their last *on they overwhelmed were day farewell by* messages and gifts

The novelty here is that word order is disrupted, but the message can still be recovered. A second experiment included a condition where scrambling was so severe that syntactic and semantic relations between words could not be established:

(5) f. last day farewell gifts on were and they by they overwhelmed message.

The results show that, if dependencies between words can be recovered, linear order has little impact on processing: activation levels were similar across conditions, irrespective of scrambling. The exception is (5f), where scrambling was such that words cannot form dependencies: here activation levels were lower, closer to the level of content word lists.

This study is a reminder of the importance of carefully constructed baseline and control conditions. When scrambled words are linearly close to other words with which they can plausibly enter a dependency relation, there are no differences between the baseline and compositional conditions. One possibility, compatible with Mollica et al.'s interpretation of these results, is that scrambled sentences (5b–e) cannot prevent extraction of meaning from input: the brain is quite "aggressive" in its urge to compose. There is another lesson one could draw here. The extent to which intepretation requires (hierarchical) syntactic structure is open to question (Culicover and Jackendoff, 2006; Baggio, 2018, 2021; Nefdt and Baggio, 2023). Participants might use linear order as a proxy for syntactic structure,[1] or extract meaning without (fully) reconstructing structure. If participants seek to compose meaning even in the scrambled conditions, either they do not need syntactic structure to compose or they are trying to fix the disrupted mapping between syntax and word order.

1 Note that linear order feeds in a systematic way off of structure, but it is not completely determined by it. Different languages have different base orders and tend to allow for variation in word order for the same message. Thus, linear order is not an exact proxy for syntactic structure.

In most fMRI studies using this paradigm, activation levels are averaged over the whole sentence and are compared with the average signal from word lists. Because of the slow temporal evolution of the BOLD response, these studies cannot zoom in on syntactic or compositional processes at specific points in a sentence, but can only indirectly associate composition to regions that activate more with the presence vs. the absence of structure, a binary variable that applies to the entire stimulus (Matchin et al., 2019a). Composition, however, is a *time-sensitive process* that may not occur in the same form at each word (there may be differences for optional vs. obligatory elements, function vs. content words etc.), that may not happen at every single word (if certain constructions imply storage of material, e.g., with long-distance dependencies), and that may be revised at subsequent processing stages (Baggio et al., 2008; Baggio, 2018). A fine-grained map of composition operations, as realized in the brain, may only be obtained from measures with sufficient temporal resolution and with experimental designs that harness that resolution. M/EEG have the advantage of sampling brain activity with a millisecond resolution. Hultén et al. (2019) use MEG to compare sentences (e.g., "I like to read nice books in my spare time") to lists containing the same words as the sentences, but in a scrambled order. For every word in the sentence, they found activity around 400 ms in the left posterior temporal cortex (LPTC), left inferior frontal cortex (LIFC), and left anterior temporal lobe (LATL). Fedorenko et al. (2016) used cortical-surface EEG (ECoG) with lists and sentences. They observed a monotonic increase of gamma power over frontal and temporal areas as the sentence unfolded. For lists, this was only seen until the third word, after which activity dropped, suggesting that participants may initially attempt to process lists much as they do sentences. Their results also show increased gamma activity for word lists relative to Jabberwocky and nonword conditions, suggesting that composition might be engaged in that condition too, as constituents in word lists may still be formed. Using ECoG, Nelson et al. (2017) compared sentences vs. scrambled lists and found high gamma decreases for words closing syntactic phrases. These studies point to possible gamma-band signatures of structure building or syntax-driven composition (but see Murphy, 2020 for a different account). However, word lists do not allow researchers to exploit the superior temporal resolution of M/EEG: as word order in lists is disrupted, one cannot compare the same word across conditions at any given time point while controlling for properties of the left context. Independent improvements of this paradigm would therefore be needed to fully take advantage of better recording resolution or advanced data analysis methods.

In these experiments, lexical material is matched between the items being compared but the presence of function words may still trigger structure building attempts also in lists, as suggested by Zaccarella et al. (2017). Their meta-analysis shows function and content words in lists can activate language regions, e.g., the left IFG. Affixes and function words carry grammatical information and can therefore guide syntactic processing. In an fMRI study of the neural correlates of syntax and semantics, Friederici et al. (2000) compared spoken German sentences (6a) to word lists (6b):

(6) a. Die hungrige Katze jagt die flinke Maus.
The hungry cat chased the fast mouse.
b. Der Koch stumm Kater Geschwindigkeit doch Ehre.
The cook silent cat velocity yet honor.

They removed function words and inflectional morphology from lists and omitted verbs: German word order can make verbs within lists trigger syntactic processing. Still, their lists are considerably less diverse lexically than sentences. They reported activations for sentences relative to lists in the bilateral superior temporal gyrus (STG). This region has been associated with phonological processes. Given the differences in length or duration of words in lists (only content words, minus verbs) vs. in sentences (content and function words), it is difficult to establish whether the STG effect here is due to composition or to processing of phonological or auditory properties of stimuli. A similar concern applies to recent work, such as Branco et al. (2020), who also used lists with only content words as baselines. They find activation for sentences relative to word lists across left frontal and temporal areas, but this result may include any area sensitive to the distinction between function and content words, as opposed to combinatorial processes more specifically.

A possible approach to isolating composition would be to remove confounding variables by modifying the stimuli in a stepwise fashion. Humphries et al. (2005) compare spoken sentences (7a) to unstructured lists with and without prosodic cues. The lists served as a baseline and could contain function and content words (7b) or only content words (7c):

(7) a. The man was looking forward to an upcoming road trip in his expensive new car.
b. That the in the wearing students the blonde expensive south up waits in performing the ate.
c. Bank calm school bathtub workers home car tambourine neail waill hat beach umbrella street head.

Permuting the words within each sentence would run the risk of accidental composition: semantically related words might prompt speakers to reconstruct a meaningful message, as was noted by the authors. They thus randomly picked words from the sentence set for scrambling, keeping the stimulus length and number of syllables constant within items. The conditions were matched lexically over the entire set, but not for each item or each sentence position. The left anterior STS, toward the middle temporal gyrus (MTG), was active for sentences regardless of prosody; the left posterior STS was active for sentences with list prosody; the posterior bilateral STS showed a prosody*structure interaction.

Lists and sentences are difficult to match in all relevant respects except for composition. Law and Pylkkänen (2021) embedded lists of nouns ("lamps, dolls, guitars") into sentences (8a) or lists (8b) in an MEG study aimed at isolating correlates of syntactic composition:

(8) a. The eccentric man hoarded lamps, dolls, guitars, watches and shoes
b. Forks, pen, toilet, rodeo, lamps, dolls, guitars, wood, symbols, straps

Their results show increased activity in the left inferior frontal cortex at 250–300 ms, at 300–350 ms in the LATL, and at 330–400 ms in the left posterior temporal cortex for lists in sentences relative to lists in lists. This design affords better control over local syntactic and semantic context, and the use of bare plural

nouns may help prevent N-N compounding in lists. However, the conditions are different beyond the immediate local context: lists do not include any function words, and content words before critical words differ between conditions, which might impact processing complexity and preactivation. Additionally, as noted by the authors, a word's meaning in a sentence could differ from the same word's meaning in a list.

2.1.2. Composition beyond sentences: Structured narratives vs. scrambled sentences

Experiments using single sentences may be argued to lack the ecological validity needed to draw inferences about how compositional machinery is used in everyday life (Hasson et al., 2018). We rarely communicate in isolated utterances: the messages that we convey often span multiple sentences. Recent studies have thus used multi-sentence narratives, typically presented in the auditory modality as naturalistic speech. Narratives have been compared to lists of scrambled words from the same story, to lists of words matched in lexical variables with words in the story, or to lists of unrelated sentences (Mazoyer et al., 1993; Xu et al., 2005; Brennan and Pylkkänen, 2012, 2017; Brennan et al., 2012). Lerner et al. (2011) compared brain responses to stories in the auditory modality with scrambled versions at different levels of structure: word, sentence, and paragraph, plus a condition with the story played backward. Using structured narratives results in more ecologically valid conditions and increases the variety of expressions investigated. But these studies also use lists of words or sentences as baseline conditions, incurring the problems raised above. Further, the size of the stimuli makes it difficult to zoom in on local composition: interpretation of most words in narratives is influenced by the discourse model built up to that stage, engaging processes beyond composition (Baggio et al., 2016; Baggio, 2018).

2.1.3. Problems with lists: Interim summary

Some paradigms have tried to align experimental and baseline conditions by controlling lexical frequency, length, and word class across sentences and lists, by scrambling words from the same sentences, by combining words from different sentences in the stimulus set, by leaving out function words, or by matching local contexts while varying aspects of global contexts. Such strategies may not always achieve minimality or precise matching of conditions (Grodzinsky et al., 2021).[2]

Beyond minimality, the potential risk of accidental syntactic or semantic composition in lists always looms over the interpretability of experimental results, particularly when the words used in lists are drawn from the critical sentences and shuffled in random order. An inspection of the stimuli used in many studies reveals that phrase level dependencies can sometimes still be formed (Mollica et al., 2020). Matchin et al. (2017) too point this out as a possibility in their list condition. The task used might encourage participants to impose syntactic structure on unstructured lists (Matchin et al., 2017). Some studies use block designs as a remedy, but drawbacks can be habituation effects or the emergence of expectations and processing strategies. There are also further differences in sentences vs. lists that are rarely discussed, for example that sentences introduce more information to be encoded in memory. Lists could engage attention and control more than sentences, if there is an active effort to interpret the stimulus.

An additional level of complexity is introduced by the interaction of problems related to the choice of methods (fMRI vs. M/EEG) with challenges that arise from problems in the paradigms themselves. With respect to minimal pairs, one question is whether the effect of noise or variability from different lexical items is more dangerous than the addition of function words in non-composition baselines, or vice versa. In fMRI, where localization is the goal, it may be more appropriate to get rid of function words than to be rigid about matching words in each comparison. With M/EEG, the trade-off might go the other way, given the prominence in measured signals of preactivation and related effects of content words, which should then be matched as much as possible. Sentence-level comparisons, for example using fMRI, would work only if differences between lists and full sentences were spatially localized on a "macro" level. Even then, fMRI's lack of temporal sensitivity still largely threatens non-minimal paradigms, if the goal is to isolate basic composition: the effects of pure composition will interleave with other linguistic operations and smear out over the total fMRI signal over the course of a sentence. This problem is exacerbated with longer discourses. Our assessment of studies using lists, scrambling, or constituent chunking is summarized in Table 1. Anomalous sentences and lists with function words are, in our view, the most problematic. Lists without function words may reduce chances of accidental composition, but the resulting contrasts are less minimal compared to lists with function words and scrambled sentences. In terms of minimality and naturalness, scrambled sentences are superior to lists with function words.

2.2. The Jabberwocky alteration: Form without content

Lists of words aim to disrupt linear order and thus prevent composition. However, this type of stimulus cannot be used to dissociate meaning and grammar: sentences and lists of words differ both in structure *and* compositional semantics (Grodzinsky et al., 2021). Differences between the two conditions will then reflect both aspects of composition.

One type of design, meant to dissociate syntax from semantics, relies on baseline stimuli that are devoid of lexical meaning, but still grammatical. Jabberwocky consist of phono- and morphotactically and grammatically well-formed strings, lacking content. Structure building is assumed to proceed unimpeded, but meaning composition is blocked by the unavailability of constituent meanings. In typical Jabberwocky experiments, all content words are replaced with phonotactically licensed pseudowords, maintaining all function words and affixes ("The gar

2 We define a "minimal pair" as two conditions that only differ in the variable of interest: e.g., conditions that only differ in that one involves composition and the other does not or where the mode of composition is different. An exact matching between conditions might prove impossible at the level of the stimulus, but a close matching might still obtain if the processes in the two conditions are identical except for the one of interest. Examples of steps in this direction are discussed in Section 3.

TABLE 1 A summary of limitations associated with each of the paradigms discussed in Section 2 with a rating (low, medium, high) of how problematic we believe each limitation is for the purposes of isolating the neural correlates of meaning composition.

Limitations	Scrambled sentences	Anomalous sentences	Lists with function words	Lists without function words	Constituent chunking	Pseudoword sentences or Jabberwocky	Minimal phrases
Comparison is not minimal	Low	High	Medium	High	High	High	Medium
Risk of accidental composition	High	High	High	Low	Medium	Medium	Medium
Lack of naturalness	Medium	Medium	High	High	Medium	High	Medium
Total problematic	Medium	High	High	Medium	Medium	High	Medium

A total average rating is also assigned to each paradigm.

was swabbing the mume from atar"; Fedorenko et al., 2016). The pseudowords are usually derived by replacing phonemes in real words while making sure that the resulting pseudowords do not exist in the given language. In Jabberwocky, syntactic constituents and dependencies are thus maintained in the absence of meaning. Some studies match low-level properties of Jabberwocky to real language by controlling variables such as bigram frequency, syllable length, and phoneme length (Heim et al., 2005; Humphries et al., 2006; Branco et al., 2020). By comparing a normal sentence (e.g., "The poet will recite a verse") with a Jabberwocky version matched in syntactic structure, but not in content (e.g., "The tevill will sawl a pand"; Matchin et al., 2017, 2019a), one can reveal brain activity that reflects processes necessary to derive compositional meaning.

There are however differences in how Jabberwocky and pseudoword sentences are used across studies. Friederici et al. (2000) maintain morphological and capitalization rules of German to give Jabberwocky the "feel" of German: "Das mumpfige Fölöfel föngert das apoldige Trekon". In addition to pseudowords, Fedorenko et al. (2016) used a low-level condition with strings of "nonwords" (e.g., "Phrez cre eked picuse emto pech cre zeigely"). This condition is meant to control for low-level orthographic processing in the absence of lexical processing and composition. Sometimes pseudowords and function words are scrambled within a sentence (e.g., "rooned the sif into lif and the and the foig aurene to"). The normal sentence vs. Jabberwocky sentence contrast is used to identify the effects of compositional semantics when structure is held constant (Röder et al., 2002), while the Jabberwocky sentence vs. Jabberwocky lists contrast is used to isolate syntactic structure building in the absence of meaning (Goucha and Friederici, 2015). This is seen as a viable strategy, if the goal is to dissociate syntactic from semantic processing (Pylkkänen et al., 2011). But as with word lists, Jabberwocky and pseudowords, let alone nonwords, raise concerns about the minimality of the stimuli compared; for example, some phonological and lexical variables cannot be measured and matched between the two conditions.

The question of whether specific areas of the language network are sensitive to syntactic structure, word meanings, and their interactions is often debated in the field (Fedorenko et al., 2012, 2020; Hagoort and Indefrey, 2014). Several studies used pseudoword sentences vs. unstructured pseudoword lists to disentangle syntax and semantics in the brain (e.g., Fedorenko et al., 2016; Matchin et al., 2017). Branco et al. (2020) use pseudowords lists, lists of content words, real word sentences, pseudoword sentences, and a non-linguistic baseline with symbols matched in length and visual features to the linguistic stimuli. A similar design is used by Humphries et al. (2006), who compared conditions assumed to be minimally different in the presence or absence of syntax or semantics. In addition to normal sentences (1a), incongruent sentences (1c), and lists (1b), they used pseudoword sentences and pseudoword lists containing real function words:

(9) a. The solims on a sonting grilloted a yome and a sovir
b. Rooned the sif into lilf the and the foig aurene to

Structured stimuli were compared to lists to establish a main effect of syntax: activation differences were seen in the left anterior STS. The effect of compositional semantics was derived by comparing normal sentences to incoherent sentences: these conditions both involve lexical processing, but only normal sentences result in a meaningful proposition. This contrast revealed effects in the left inferior temporal gyrus, the left STS, and the left AG. Comparisons were performed between incoherent and pseudoword sentences (with activation in left anterior, middle, posterior STS) and between normal and pseudoword sentences to determine effects of lexical processing (anterior, middle, posterior STS and MTG). The analysis was limited to temporal areas, but the results show that semantics is subserved by a wider network of areas in the temporal lobe than syntax.

Stromswold et al. (1996) used a variation of this paradigm with conditions in which only one word in a sentence was replaced by a pseudoword (10a) vs. center-embedded (10b) and right-branching (10c) sentences:

(10) a. The economist predicted the recession that *chorried* the man
b. The limerick that the boy recited appalled the priest
c. The biographer omitted the story that insulted the queen

By manipulating both syntactic complexity and the possibility of deriving compositional meaning, this study asks whether brain areas subserving syntax as opposed to semantics can be isolated. They found increased activation in LIFG for syntactically more complex sentences and in the inferior frontal gyrus, superior temporal gyrus, and supramarginal gyrus for normal sentences vs. sentences with a pseudoword.

Another experiment using pseudowords to investigate syntactic composition is Segaert et al. (2018). To minimize the effect of

semantics, they used sentences where the subject is a pronoun and the verb is a pseudoword with inflectional morphology ("She grushes"). The baseline is a list of pseudowords matched in length to the sentences ("pob grushes"). The pronoun is assumed to trigger syntactic composition, whereas the pseudowords list should not. Structure building could also occur in lists, as morphological marking on the second word could allow speakers to parse the list as a pseudo-subject noun followed by a pseudo-verb. The study found increases in EEG alpha power over left fronto-temporal channels for sentences vs. lists, for the first and second words, interpreted as predictive and syntactic processes respectively (see also Hardy et al., 2023).

It could be argued that Jabberwocky still involves formal compositional semantics, even though lexical and conceptual semantics are absent. Grammatical cues could license the assignment of thematic roles toward an interpretation: e.g., "The tevill will sawl a pand" refers to an event (sawl) that will be initiated by an entity (the tevil) affecting another (a pand). This is compatible with the results of studies such as Branco et al. (2020),[3] which did not find differences in activation between real sentences and pseudoword sentences. Goucha and Friederici (2015) exemplify this observation in a parametric design. To identify areas of the left inferior frontal gyrus selectively involved in syntax and semantics, they used several types of pseudoword sentences as baselines. Their Jabberwocky sentences contained phonologically licensed pseudo-content words and real function words, with inflectional and derivational morphology (10a). They removed derivational morphemes (10b) and inflectional morphology replacing determiners with pseudowords (10c):

(10) a. The pandex*ity* of the larisa*tion***s** ha**d** zapp**ed** the *un*heggy wog*dom*.
b. The pandesteek of the larisarden**s** ha**d** zapp**ed** the enhegg**ed** fordem.
c. Thue pandesteek of thue larisarden feg zopp thue enheg fordem.

Their fMRI results show a different pattern of activation for pseudoword sentences with vs. without derivational morphology, suggestive of residual morphosyntactic processing.

Another known issue is that pseudowords, due to their resemblance to real words, might trigger a "search" in the lexicon which will return no results. This might make them more difficult to process than real words, undermining the assumption that pseudowords can serve as a baseline involving fewer/simpler processes. Iwabuchi and Makuuchi (2021) use pronounceable letter strings as placeholders for real words, adding relevant morphology to form hierarchical structures in Japanese. They also included a syntactic manipulation with sentences with the canonical SOV word order (11a), more complex OSV order (11c), as well as non-semantic sentences containing placeholders, but with the same syntactic structures as the natural sentences (11b, d):

(11) a. ranboo-na sootoku-ga daijin-o tataita. (*The wild governor hit the minister.*)
b. PP-na AA-ga BB-o V-sita. ($PP_{adjective}$ *-AA V-PAST BB*)
c. daijin-o ranboo-na sootoku-ga tataita. (*The wild governor hit the minister.*)
d. BB-o PP-na AA-ga V-sita. ($PP_{adjective}$ *-AA V-PAST BB*)

This type of design aims at dissociating syntactic from semantic processes in the brain, without using an additional condition of pseudowords and word lists. Using fMRI, they found an effect in the LATL for sentences vs. pronounceable non-sentences regardless of word order. BA44, premotor, and parietal cortices were more active to the placeholders. This latter finding might be attributed to the perceptual and/or phonological differences between placeholders and real words. The effect of syntax was less robust: activations in BA45 and pMTG were observed only before correcting for multiple comparisons.

2.2.1. Problems with Jabberwocky: Interim summary

The Jabberwocky paradigm tries to create an impoverished language, where meaning is removed but syntactic structure is preserved: the goal is to block semantic composition while keeping syntactic composition and other grammatical processes going. However, pseudoword sentences do not entirely lack compositional meaning, and function words, when present, can trigger the construction of a minimal formal semantic representation. Comparing sentences to Jabberwocky, with the purpose of isolating processes specific to meaning composition, can result in loss of signal precisely relevant to the latter process. Pseudowords and real words differ in frequency, familiarity, and the cognitive resources allocated to them, for example lexical recognition and search. Pseudoword sentences are used as part of designs also including (pseudo-)word lists, but pseudowords and lists of real words differ in their levels of salience and intelligibility, making direct comparisons difficult (Bautista and Wilson, 2016). Studies attempting to isolate syntactic and semantic components of language processing using word lists and pseudowords sentences can fail to create true minimal pairs: these conditions differ on other dimensions from sentences than just the presence or absence of syntax and semantics (Grodzinsky et al., 2021). Our assessment of studies using pseudowords sentences or Jabberwocky is provided in Table 1. Lack of minimality and naturalness of these stimuli are the main limitations and what renders these paradigms overall problematic for studying meaning composition.

2.3. Minimal phrases

Sentences involve processes that can obscure purely compositional operations. Semantic associations and other memory-based processes, conceptual combination, preactivation, prediction, and inferential, referential, and elaborative processes, among others (Baggio, 2018), contribute to meaning construction over and above composition. These processes interact with each other to ease demands on processing of downstream inputs (Bemis and Pylkkänen, 2013a; Zaccarella et al., 2017). In none of the paradigms reviewed above can composition be fully disentangled from co-occurring processes. Previous sentence-level studies have focused on delineating linguistic distinctions, such as lexicon

3 This is also the explanation given by the authors for the lack of an effect.

vs. grammar, under the assumption of large-scale differences in localization. Interpreting their results to make claims about Compositionality requires linking hypotheses on the role of syntax in composition, e.g., whether syntax is the only driver vs. one constraint among many, or whether composition differs for lexical content vs. logical syntacto-semantic relations.

In order for a compositional algorithm to be set in motion, it needs to be fed at least two elements (e.g., words) to produce the meaning of their combination. From a generativist standpoint, elements are combined in pairs. This combination then becomes an element too, to be combined with another in a further step of the derivation. The minimal phrase paradigm, by Pylkkänen and collaborators, uses two-word phrases as the main object of investigation. Bemis and Pylkkänen (2011) "truncate" the pseudowords and lists designs in order to adapt them for the study of composition in simple phrases. Their compositional stimulus was a two-word uninflected adjective-noun phrase ("red boat") to be compared to a baseline consisting of an unpronounceable letter string followed by the same noun ("xkp boat"). The noun "boat", at which the comparison is made, can enter composition in the first but not in the second condition. The use of an unpronounceable letter string, as opposed to a pseudoword, would serve to prevent composition attempts. To control for influences of the lexical material before "boat" in the two word conditions, they included non-combinatorial lists of two nouns ("cup, boat"). However, the brain is eager to extract meaning from input, and there is a possibility of noun-noun compounding in lists (e.g., a plastic or paper cup made to float like a boat). Bemis and Pylkkänen then introduce an additional task manipulation. The task required participants to compose the meaning of the two words and to check whether the combination matched a subsequent picture of a colored object (composition task) vs. read each word to verify whether one matches the picture following each trial (non-composition task). Composition only takes place at the second word, where contextual processes are minimized. This makes minimal phrases a better fit for time sensitive M/EEG methodology than other paradigms. In the auditory modality, pink noise can be used as a baseline instead of nonwords (Bemis and Pylkkänen, 2013b). Activity in the LATL, from around 200 ms from the onset of the second word, has emerged as a possible signature of semantic combination (Pylkkänen, 2019).

This paradigm combines a tightly controlled stimulus set with manipulations of the task to ensure that the recorded brain activity is related to the process at issue. For example, Bemis and Pylkkänen (2013a) compare canonical adjective noun phrases ("red boat") with reversed counterparts ("boat red") and nonword-word strings ("xhl cup", "frw red"). The key manipulation is the task, which involves a colored shape (compose) or two pictures, one of a colorless shape and one of a colored blob (non-compose): participants had to respond whether the probe matched both words. This study tested whether composition can also be deployed in ungrammatical sequences and whether it is automatic enough to be engaged even when the task does not require it. They found that the LATL is engaged in reversed sequences only when the task requires composition and with canonical word order regardless of the task. Fló et al. (2020) show that, when the task manipulation is eliminated, the effects of composition are no longer observed with EEG.

The minimal phrase experiments achieve something which has been challenging for the previously discussed paradigms: matching between conditions the word which has to be composed or not, at the position at which the neural signal is measured. The pre-critical content in non-combinatorial conditions (nonwords and nouns in lists), however, differs in several respects from the adjective used in the compose conditions. These differences might affect the signal recorded at the critical word. For example, a nonword at the start of a trial might make participants less engaged in processing the following words. At the same time, preactivations resulting from processing of a noun in lists and of an adjective in compositional trials will differ. Additionally, the two word list condition might trigger a process of compounding and thus involve composition regardless of explicit task.

Some minimal phrase studies have used multiple and different baseline conditions. Neufeld et al. (2016), Fritz and Baggio (2020, 2022), and Kochari et al. (2021) use pseudowords and nonwords to disentangle semantic and syntactic processes, and Bemis and Pylkkänen (2013a) use a reversed word order condition ("boat red"). Del Prato and Pylkkänen (2014), instead of lists of nouns, use lists of adjectives and lists of numerals as baselines, which match in category to the precritical words used in the combinatorial contexts. Graessner et al. (2021a,b) contrast meaningful two-word phrases ("fresh apple") to anomalous phrases ("awake apple") and adjective-pseudoword phrases ("fresh gufel"). In an ECoG experiment, Murphy et al. (2022) compare adjective-noun phrases ("red boat"), which are assumed to involve composition at the noun and prediction at the adjective, to adjective-pseudoword phrases ("red neub"), involving just prediction, and to pseudoword-noun phrases ("zuik boat"), which involve neither.

Some minimal phrase studies have tested how different semantic contexts interact with composition, for example how specificity of the noun modulates LATL activity (Zhang and Pylkkänen, 2015) and the impact of semantic properties of adjectives (e.g., see Ziegler and Pylkkänen, 2016; Fritz and Baggio, 2020, 2022; Kochari et al., 2021). Kim and Pylkkänen (2019) look for MEG correlates of composition in adverb-verb constructions, testing whether different classes of adverbs (eventive "slowly" vs. orientative "reluctantly") show similar LATL effects as in adjective-noun phrases. Manipulations of the precritical word target the interplay of composition and prediction, *via* the use of different pronoun types (Strijkers et al., 2019), and between composition and semantic properties of nouns, such as relationality or eventivity (Boylan et al., 2017; Williams et al., 2017). Studies have revealed early LATL responses for Adj-N phrases in the auditory and visual modalities. However, Kochari et al. (2021) failed to replicate this finding. The sensitivity of LATL to variables that syntax-driven composition should, according to theory, *not* be sensitive to (e.g., specificity) has led to the conclusion that the LATL does not perform composition, but rather *conceptual combination* (Pylkkänen, 2019). Moreover, the angular gyrus (AG) and the ventromedial prefrontal cortex (vmPFC) are involved in semantics, though they do not always activate across studies. Murphy et al. (2022) find effects of composition in

portions of the pSTS using iEEG/ECoG. With EEG and minimal phrases, Neufeld et al. (2016) link the N400 to combinatorial semantic processing (Hagoort et al., 2009; Baggio and Hagoort, 2011; Baggio, 2012; Nieuwland et al., 2020), and Fritz and Baggio (2020, 2022) find and replicate P600 effects for adjective-noun composition.

The relatively tight control over experimental items offered by minimal phrases has also been used to tackle more fine-grained and theoretically relevant questions on the nature of composition. One question is whether composition in different syntactic structures or environments, such as modification and predication, is carried out by different neural processes. Westerlund et al. (2015) test the distinction between composition operations of *argument saturation* and *predicate modification* (Heim and Kratzer, 1998): the former mode of composition includes verb-noun (e.g., "eats meat"), preposition-noun ("in Italy"), and determiner-noun ("Tarzan's vine") combinations; the latter includes adjective-noun (e.g., "black sweater"), adverb-verb ("never jogged"), and adverb-adjective ("very soft"). In keeping with the standard design, each expression was compared to a nonword followed by a matched noun in order to establish effects of composition. Boylan et al. (2015) use a similar design, crossing mode of composition (argument type: "eats meat", "with meat" or adjunct type: "eats slowly", "tasty meat") with presence or absence of a verb. The baseline was non-compositional phrases in which the nonword was either the first or the second element of the sequence ("eats fghjl"/"fghjl eats"). A similar approach is used by Schell et al. (2017). Matchin et al. (2019b) matched word forms exactly within the phrases, while varying syntactic structure for noun-adjective (e.g., "the frightened boy") and verb-noun ("frightened the boy") composition. A potential confound might arise in these designs, as also noted by Matchin et al. (2019a). Whereas, a noun composed with a modifier may be interpreted as a saturated structure on its own, a noun in the object position, composing with a verb, results in incomplete syntactic and semantic structures. Boylan et al. (2015) report activity in the left AG, regardless of mode of composition, for "eats meat" vs. "tasty meat". Westerlund et al. (2015) found that the LATL is involved in argument saturation *and* predicate modification. Matchin et al. (2019b) show that activity in the left IFG and pSTS increases for verb-noun composition, while there is no difference between the two syntactic structures in AG and LATL activation.

It is worth mentioning two more studies that extend the minimal phrase paradigm. Kim and Pylkkänen (2021) use hashtags in various positions in sentences to study subject-verb composition vs. verb-object composition (e.g., "kids toss objects" vs. "### toss objects" vs. "### ### objects"). However, hashtags can discourage participants to compose meaning for the rest of the sentence, as noted by the authors. Lau and Liao (2018) used coordinated adjective-noun phrases (e.g., "sunlit ponds and green umbrellas") vs. those noun phrases separated by hashtags ("sunlit ponds ### green umbrellas") vs. Jabberwocky versions to isolate brain correlates of building coordinated structures. They find sustained anterior negative ERPs from the first word in the second phrase for coordinated constructions.

2.3.1. Problems with minimal phrases: Interim summary

The elegance and simplicity of the minimal phrase paradigm has provided fertile ground for testing core linguistic ideas with M/EEG. The main advantage of this paradigm is the control it affords over experimental stimuli, enabling the minimization of processes not strictly reflecting local combinatorics. However, minimality comes at a cost, for example a loss of naturalness or ecological validity of stimuli (Hasson et al., 2018). Full sentences may not be the most frequent type of utterance in *spoken* language corpora, but neither are NPs or VPs as used in these experiments; when those occur, they are elliptic phrases, interpretable in the context of other utterances. Most of these experiments used *written* stimuli: in written corpora disconnected noun or verb phrases may be even less common than in spoken corpora. However, one could argue that composition must take place for any given phrase, regardless of whether a naturalistic context is available. Another issue is that the baselines used in these experiments may differ from phrases in other respects than just composition. Our assessment of the minimal phrase paradigm is given in Table 1. This paradigm compares favorably to many others currently in use and is the one with the best balance between different limitations.

3. Alternative and emerging approaches

3.1. Theory-inspired and language-specific manipulations

For the paradigms just discussed, linguistic theory only covers combinatorial conditions, and possibly Jabberwocky and semantically anomalous sentences, but offers no analysis of conditions with lists of words, pseudowords, nonwords, and scrambled sentences. To bridge levels of analysis with linking hypotheses that can be evaluated empirically, both combinatorial and baseline conditions should be covered by formal theories: ideally, our theories should state why and how composition applies to some cases but not to others.

To design experiments capable of addressing composition, theoretical distinctions must be identified in the linguistics literature and stimuli reflecting those distinctions must be constructed. Consider complement coercion (Pylkkänen, 2008). Semantically, aspectual verbs, such as "begin" and "finish", require event-denoting complements (e.g., "begin the fight"), but syntactically, they may be combined with entity-denoting complements (e.g., "begin the book"): the denotation of the NP must then be coerced from entity to event, or an equivalent (e.g., inferential) operation must recover an eventive interpretation of the NP. In coercion constructions, syntactic structure is simple, but composition load varies: it is greater for entity-denoting than for event-denoting NPs (Piñango and Deo, 2016).

Baggio et al. (2010) and Kuperberg et al. (2010) compared control conditions (12a) with coercion constructions (12b) and semantic anomalies (12c). Similar conditions were also used by Pylkkänen and McElree (2007) and Husband et al. (2011):

(12) a. The journalist wrote the article
b. The journalist began the article
c. The journalist astonished the article

These studies did not use non-combinatorial baseline conditions that attempt to prevent composition, but vary processing load between two conditions that require composition, while keeping plausibility and semantic associations from the context before the critical noun ("article") as constant as possible. This strategy has also been applied to metonymic constructions (Schumacher, 2013) and aspectual coercion (Paczynski et al., 2014). Baggio et al. (2010) and Kuperberg et al. (2010) find N400-type ERP negativities. Using MEG, Pylkkänen and McElree (2007) find increased activation of vmPFC for coercing sentences. Schumacher (2013) reports late positivities for container-for-content metonymies (e.g., "The baby drank the bottle"). Paczynski et al. (2014) demonstrate that aspectual coercion (i.e., composition of punctual verbs and durative adverbs, e.g., "For several minutes, the cat pounced on the toy") is indexed by a late anterior negative ERP. In these studies, the conditions are closely matched, but precritical material is not kept constant. The focus on semantic differences between conditions, motivated by theory, is a valid way forward to investigate the online processing of these constructions and has the potential to refine linguistic theories. Still, the variable results emerging from these studies point to effects specific to the different linguistic phenomena investigated by each study as opposed to a neural correlate unique to composition.

Other studies are designed around syntactic or semantic properties of languages. Flick and Pylkkänen (2020) use properties of English in an attempt to vary syntax while keeping meaning constant. In English, attributive adjectives occur canonically before a noun, but they may also occur post-nominally in specific constructions. They compared declarative sentences with post-nominal modifiers ("There are many trails wide enough for a bear") to questions with post-nominal predicative adjectives ("Are many trails wide?"). A novel aspect here, which is not found in minimal phrases, and to which we return later, is that the critical and pre-critical words form *identical sequences across conditions* ("... trails wide …"). The authors find an effect of structure in the left posterior temporal lobe (PTL) around 200 ms after the onset of the adjective, and an effect of semantic fit between the adjective and noun in the LATL.

Parrish and Pylkkänen (2022) use semantic and syntactic properties of English to vary the point of composition. They compare expressions where an adverb and an adjective enter into local composition (e.g., "pleasantly sunny days") to expressions where two adjectives compose with a noun, but not locally with each other (e.g., "pleasant sunny days"). In this study, the precritical word was matched across conditions at the lemma or concept level, but not in its grammatical form. A further comparison involved structures such as "this herbal tea", where "tea" and "herbal" readily combine with each other, to conditions where they do not because of a gender mismatch: "these herbal tea …". In this case, participants must wait until they see a noun that closes the phrase, like "these herbal tea drinkers". A non-combinatorial condition was created by placing the critical word at the start of the sentence, where it has no previous material to combine with: "Tea drinkers hate coffee". Composition in LATL can proceed in the absence of syntactic phrase closure, but syntax can also influence activity in this region, with the highest activity seen for phrases that were both syntactically and conceptually straightforwardly composable.

Matchin et al. (2019b) exploit the fact English participle adjectives and past tensed verbs have the same form to construct modification and predication pairs (e.g., "the frightened boy" vs. "frightened the boy"), plus a list baseline (e.g., "frightened, scrubbed, wounded"). They found no differences in BOLD responses in the left ATL and AG. The left posterior STS and LIFG showed greater activity for predication (VP) vs. modification (NP). Matar et al. (2021) use unique properties of the Arabic language to achieve minimally differing stimuli where only syntactic composition varies. In Arabic, an adjective follows the noun it modifies. If the adjective and noun carry the definiteness marker (e.g., "al", in "al-kursi al-banafsaji", the purple chair) the result is an NP; if only the noun does (e.g., "al-kursi banafsaji", the chair is purple), a full sentence results. These two conditions were further compared to an indefinite NP (e.g., "kursi banafsaji", a purple chair). There were no MEG effects of syntactic structure in the left IFG, ATL, and AG. The left posterior temporal lobe (LPTL) was engaged more for indefinite NPs than for definite NPs, and least of all for sentences. The direction of this effect (NP > S) is opposite to that reported by Matchin et al. (2019b) (VP > NP) in the same region of the left posterior temporal cortex. Using a similar approach, Artoni et al. (2020) used Italian sentences containing noun phrases or verb phrases containing homophone two-word sequences, e.g., "la porta" in (13), which is either a Det-N phrase (13a) or a clitic followed by a verb in (13b) (the fragment "domani la porta" is in fact structurally ambiguous: Adv-VP vs. Adv-NP):

(13) a. Pulisce **la porta** con l'acqua.
*[He/she] washes **the door** with water.*
b. Domani **la porta** a casa.
*[He/she] tomorrow **takes her/it** at home.*

Using direct cortical EEG recordings, they found increased gamma activity above 150 Hz for VPs compared to NPs in large portions of the left hemisphere, beyond the LIFG and posterior STG/STS. The studies presented in this section compare conditions where the degree or type of composition varies to identify correlates responsible for the difference. However, to isolate composition true non-combinatorial conditions that do not have the limitations discussed so far would be needed.

3.2. Frequency tagging paradigms

Another approach to the study of structure building and indirectly meaning composition is the frequency tagging (or neural tracking) paradigm. By using rhythmically presented stimuli, recent studies have shown that neural oscillations in particular frequency bands can align with chunks at different levels of syntactic structure, as shown by peaks in the power spectrum of particular frequency bands (Ding et al., 2016) or increases in mutual information (MI) between auditory stimuli and neural oscillations (Kaufeld et al., 2020).

Ding et al. (2016) and Sheng et al. (2019) compared scrambled syllable sequences with 4-syllable sentences and 4-syllable NPs

and VPs, matched in length but differing in the point at which structural dependencies are formed. They found rhythmic brain activity tracking each level of structure: syllable, phrase, sentence. There were no prosodic cues or breaks between sentences in a sequence: those effects can be attributed to synchrony of neural activity to internally generated structures (Meyer et al., 2020; see Kazanina and Tavano, 2023 for discussion). While Sheng et al. (2019) use MEG, Ding et al. (2016) also present ECoG data. They found activity modulated at the phrase frequency in bilateral pSTG, and in the left IFG and pSTG at the sentence frequency.

Coopmans et al. (2022) compared normal sentences (14a) to idiomatic sentences (14b), anomalous prose (14c), Jabberwocky (14d), and scrambled sentences (14e):

(14) a. De jongen gaat zijn zusje met haar huiswerk helpen.
The boy will help his sister with her homework.
b. De directie zal een vinger aan de pols houden.
The directorate will keep a finger on the wrist.
c. Een prestatie zal het concept naar de mouwen leiden.
An achievement will lead the concept to the sleeves.
d. De jormen gaat zijn lumse met haar luisberk malpen.
The jormen will malp his lumse with her luisberk.
c. De gaat jongen zusje huiswerk zijn haar helpen met
The will boy sister homework his her help with

This study shows how a combination of different baseline conditions and advanced data analysis techniques allows us to track neural dynamics across conditions. At the phrase frequency, there were no differences in MI between sentences and anomalous prose, or sentences and idioms, but they found increased neural tracking in sentences compared to lists and Jabberwocky, as in Kaufeld et al. (2020). ERPs show differences between all of these conditions, but neural tracking reveals similarities across conditions containing structure and content words, pointing to a common mechanism for composition.

Burroughs et al. (2021) adapt the paradigm used by Ding et al., in an experiment aimed at disentangling the effects of word category repetition from those of structure building. They found that the neural signal tracks syntactic structure, with increased tracking in the delta band for lists of phrases ("cold food loud room tall girl") vs. lists of words with repetitions of syntactic categories without structure ("rough give ill tell thin chew"). This effect is however modulated by syntactic category, with reduced tracking when the list of phrases did not contain repetition of syntactic categories ("that word send less too loud"). These results suggest that previous studies using the frequency tagging paradigm may have also included spurious effects of syntactic category repetition.

Glushko et al. (2022) use EEG to disentangle the effects of syntax from those of prosody. They used sentences containing four words of the form NP-VP, with the NP consisting of 1 word (1+3 Syntax) or 2 words (2+2 Syntax) without prosody. These were then compared to trials containing the same syntactic structures but with a prosodic contour compatible with the 2+2 Syntax condition. Their results show an interaction between prosody and syntactic structure, suggesting that the generation of implicit prosody affects syntactic composition and that previously reported effects using the neural tracking paradigm can be partially explained by prosody effects.

Kalenkovich et al. (2022) used Russian sentences containing the same number of words and lexical content and differing only by the use of a single suffix, which affords them a different syntactic structure. They created sentences with words grouped into 2 phrases (Genitive 2-2) and sentences containing a noun in the dative case with the same words grouped in a 1 word NP and a 3 word verb phrase (Dative 1-3). Interestingly, the spectral peaks between conditions at the 2-word frequency did not differ, suggesting that factors like repetition of lexical category might explain previous effects.

The frequency tagging paradigm has become popular since its introduction by Ding and colleagues. The conclusions originally drawn from those experiments have been recently challenged on empirical and theoretical grounds (Kazanina and Tavano, 2023), suggesting that the rhythmicity of stimulus presentation may introduce processes that stand in the way of observing neural correlates of structure building.

3.3. The cut-compose paradigm

The studies reviewed in Sections 1–2 investigate composition by comparing well-formed language to baselines that are assumed not to engage composition. It is unclear to what extent pseudowords and word lists prevent composition: composition-related signal can be lost if both conditions under comparison engage composition. A second challenge is that those baselines can differ from compositional expressions on several levels besides composition, leaving in mixed signals after subtraction or comparison. A third difficulty is that pseudoword sentences, word lists, and phrases are not as natural and informative as full sentences and can require additional pragmatic support, when they do not violate pragmatic constraints altogether.

We describe a novel paradigm for studying composition which tries to take into account the three limitations of previous paradigms: lack of minimality, lack of naturalness, and unsuccessful prevention of composition. The goal here is to learn from the successes and failures of previous studies and to explore possible new avenues in experimental design.

The Cut-Compose paradigm makes use of natural, well-formed, and complete sentences, varying the presence or absence of composition at specific points in the input string. The idea is to force or prevent composition in well-formed, meaningful sentences or pairs of sentences by exploiting syntactic boundaries:

(15) a. Some birds sit on [***grey elephants***] and clean them.
b. Some birds are completely [***grey***.][***Elephants***] can be white.

The same sequence of two words can occur as part of the same constituent, in (15a), the Compose condition, or as separated by a syntactic boundary, in (15b), the Cut condition, in this case also marked by punctuation. The first EEG study using this design, by Olstad et al. (2020), removed punctuation marks in order to match the precritical (e.g., "grey") and critical ("elephants") words. Additional safeguards had to be implemented to prevent accidental composition in the Cut condition. First, syntactically, the adjective "grey" has a predicative role, so it cannot modify "elephants".

Second, semantically, "Some birds are completely grey elephants" would be anomalous. Third, the critical word initiates a new sentence, rather than a new phrase in the same sentence; this should block composition of larger constituents (e.g., phrases or clauses) higher up in the syntactic structure. One challenge is to match the precritical context in length, grammatical complexity (e.g., in syntactic nodes or arcs) and semantic associations: this is crucial for experiments using hemodynamic methods, while M/EEG studies should also attempt to control the factors that affect composition locally, around the boundary. The difference between Compose and Cut is meant to reveal that which differs between the two conditions, namely the composition of the adjective "grey" with the noun "elephants" in (15a) but not (15b).

Similar to other paradigms, Cut-Compose also affords the possibility of investigating the compositional mechanisms involved in different semantic and syntactic contexts. Olstad et al. (2020) compared modification as in (15), with predication constructions as in (16), to assess whether these two different "modes of composition"—Predicate Modification vs. Functional Application, Adjoin vs. Merge—correspond to different neural events. As the study was conducted in Norwegian, the Cut sentence was created by fronting the object:

(16) a. bråk er slitsomt men noen [**hører musikk**] blant alle lydene
noise is tiring but some [*hear music*] *among all the sounds*
b. bråk er innimellom noe man **hører musikk** er flott
noise is sometimes something one hears music is nice

In (16a), the proposition is incomplete without "musikk", as the verb "hører" requires two arguments to be saturated. This contrasts with Cut (16b), where the verb argument slots are all filled by "hører", leaving no room for "musikk" to compose with the verb. Different modes of composition can be directly compared in the same experiment, as the noun at which the M/EEG signal is measured can be held constant across environments. The sentences in (17) are examples of stimuli in the modification condition Olstad et al. (2020):

(17) a. på byggeplasser spilles [bråkete musikk] på radioen
on construction sites is played [*noisy music*] *on the radio*
b. byggeplasser er bråkete][musikk kan være avslappende.
construction sites are noisy][*music can be relaxing*

Olstad et al. (2020) found different ERP signals for the different modes of composition, providing support for the theoretical distinction between predication and modification, as well as preliminary evidence for the viability of the Cut-Compose paradigm.

Does composition not happen at all in the Cut condition? In both conditions, the critical noun is eventually composed into a higher-order representation: it is combined with the previous words in Compose, while it is yet to be combined with subsequent material in Cut. However, in the Cut condition, composition does not occur *between the noun and its preceding context*, and this the key difference with Compose. In contrast to artificial stimuli such as nonword or pseudoword strings, in Cut/Compose participants should be equally engaged in reading both types of sentences, implying a more equal distribution of cognitive resources (attention, memory etc.) across conditions. Additionally, both Cut and Compose are covered by theory: all formal linguistic theories on the market predict that composition is triggered in one case but not the other, at the point of measurement.

As other paradigms, Cut/Compose has limitations related to the baseline condition. One potential issue is the use of punctuation, which is necessary in order to make the stimuli as natural and as unambiguous as possible. Adding a period after the precritical word in Cut sentences creates a perceptual difference between the two conditions. An additional perceptual difference is capitalization of the first letter of the critical word in Cut. Olstad et al. (2020) avoided the use of punctuation and capitalization, relying on the structural properties of sentences to ensure that the noun is interpreted as starting a new sentence in the Cut condition. Follow-up experiments are needed to investigate the effects of both punctuation and capitalization in the visual modality, whether they affect the detection and quantification of composition signals, and the corresponding impact of appropriate prosody or intonation around the Cut boundary in the auditory modality.

Another possible issue is that critical nouns in the Cut condition introduce a new phrase and sentence, and may therefore engage different processes than nouns in the Compose condition which *close* a phrase or sentence. This issue may be partly addressed in future experiments where the syntactic cut is not a sentential boundary but a phrasal one. Note that inferences drawn regarding different modes of composition should still be valid, as opening a new sentence in the Cut condition should involve the same processes for both predication and modification contrasts. A different issue is that of discourse processing. The second sentence in the Cut condition is not disconnected from the first one. At the critical noun, the participant might try to integrate it into the discourse model instead of waiting to read the rest of the second sentence. However, integration with the preceding context also happens in Compose sentences, though the discourse representation in that case is not organized into multiple sentential or propositional units.

Similar to constituent chunking studies, like Pallier et al. (2011), Cut/Compose relies on manipulating the number of syntactic units between conditions, while it tries to control more precisely the immediate context of the critical word as well as aspects of the wider semantic context. Cut-Compose can be used with a variety of constructions, differing in semantic or syntactic properties, complexity and length. Many questions that have been of interest for other paradigms can also be tested with Cut/Compose: coercion, different classes of adjectives, adverbs, nouns and verbs, as well as the composition of functional and lexical elements. In the long run, we will be able to inch closer to the mechanisms by which the brain builds structure and meaning only by integrating results from different paradigms, different measures and data analysis methods. Cut/Compose aims to make a contribution to this longer-term project, and might also prompt the development of new and improved paradigms beyond the currently available ones.

10.3389/flang.2023.1096110

TABLE 2 Summary of designs and results from a selection experiments on syntactic structure building and semantic composition grouped according to the paradigm used.

Paradigm	References	Results	Task	Acquisition method	Stimuli presentation
Normal sentence vs. scrambled sentence	Kaufeld et al., 2020	Increased neural tracking at phrase frequency in sentences vs. scrambled sentences	No task	EEG block design	Auditory
	Vandenberghe et al., 2002	Left anterior temporal pole, left anterior STS, left posterior temporal gyrus	Press a button if two stimuli followed each other	PET block design	Visual
	Humphries et al., 2006	Left anterior STS, left inferior temporal gyrus, left AG, left ATL	Rate stimuli for meaningfulness	fMRI event-related design	Auditory
	Hultén et al., 2019	Left PTC, left IFC, left ATL 400 ms after word onset	Yes/no question (20% of trials), word probe task for lists, comprehension question for sentences	MEG block design	Visual
	Mollica et al., 2020	Left IFG, left ATL, left PTL, left MFG, left AG for intact or moderately scrambled items vs. fully scrambled items	Word probe task after each trial	fMRI event-related design	Visual
	Nelson et al., 2017	High-gamma power increases at each new word, decreases when a word completes a phrase: left temporal, inferior frontal cortex	Sentences probe task in sentence trials, word probe task in lists trials (75% of trials)	ECoG event-related design	Visual
Normal sentence vs. anomalous or incongruent sentence	Vandenberghe et al., 2002	No effect; effect of anomalous vs. normal sentence in left MTG	Press a button if two stimuli followed each other	PET block design	Visual
	Humphries et al., 2006	Left AG, ITS, ITG, anterior STS	Rate each stimulus for meaningfulness	fMRI event-related design	Auditory
Normal sentence vs. word lists (without function words)	Friederici et al., 2000	Left posterior STG, planum polare bilaterally	Indicate whether a target word or syntactic structure was present in the previous trial	fMRI event-related design	Auditory
	Branco et al., 2020	Left IFG; left TP, MTG, SMG; left SFG, MFG; right STG; right TP	Word probe task: select which of two words was present in the previous trial	fMRI block design	Visual
	Law and Pylkkänen, 2021	Left IFG (250–300 ms), left ATL (300–350 ms), left PTC (330–400 ms)	Word probe task	MEG event-related design	Visual
	Zaccarella et al., 2017	Left IFG, left posterior STS	Decide whether the previous trial was a phrase/sentence or word list	fMRI block design	Visual
Normal sentence vs. word lists (with function words)	Humphries et al., 2005	Left posterior STS, left anterior STS/MTG	No task	fMRI event-related design	Auditory
	Fedorenko et al., 2016	Gamma increase in left frontal, left lateral temporal, left ventral temporal cortex	Word probe task	ECoG event-related design	Visual
	Matchin et al., 2017	Left IFG, STS, ATL	Word probe task	fMRI block design	Visual
	Pallier et al., 2011	Increased activity with constituent size in left IFG, TP, TPJ, STS	Rare probe sentence asking to press a button on the basis of previous trial and a word memory test at the end of each run.	fMRI event-related design	Visual

(Continued)

TABLE 2 (Continued)

Paradigm	References	Results	Task	Acquisition method	Stimuli presentation
Constituent chunking: sentence vs. phrase	Shain et al., 2021	Left IFG, MFG, ATL, PTL, AG	No task?	fMRI event-related design	Visual
	Matchin et al., 2017	Left IFG, posterior STS, ATL	Word probe task	fMRI block design	Visual
	Matchin et al., 2019a	Left ATL, left PTL (subject NP), left TPJ (object NP)	Word probe task	MEG block design	Visual
Normal sentence vs. pseudoword- or nonword-sentence	Friederici et al., 2000	No effect	Indicate whether a target word or syntactic structure was present in the previous trial	fMRI event-related design	Auditory
	Branco et al., 2020	No effect	Word probe task: select which of two words was present in the previous trial	fMRI block design	Visual
	Fedorenko et al., 2016	Gamma increase in left frontal, left lateral temporal, left ventral temporal cortex	Word probe task	ECoG event-related design	Visual
	Humphries et al., 2006	Anterior, middle, posterior STS, MTG, left ITG, bilateral AG	Rate each stimulus for meaningfulness	fMRI event-related design	Auditory
	Stromswold et al., 1996	Left IFG, left STS; left SMG gyrus (for reverse contrast)	Judge the goodness of each sentence	PET block design	visual
	Segaert et al., 2018	Alpha and beta power increases after presentation of first word; alpha power increases after presentation of second word immediately after word onset	Detect reversed speech segments	EEG event-related design	Auditory
	Iwabuchi and Makuuchi, 2021	Left ATL, ventral occipital cortex (placeholders instead of pseudowords)	Judge whether the content of a probe sentence matched the content of the previous trial (task after 60% trials)	fMRI event-related design	Visual
	Kaufeld et al., 2020	Increased neural tracking at phrase frequency	No task	EEG block design	Auditory
	Matchin et al., 2017	Left IFG, left ATL, left PTL (whole brain analysis)	Word probe task	fMRI block design	Visual
	Matchin et al., 2019b	Left IFG, left ATL, left PTL, left TPJ (all at 215–350 ms after open class word onset)	Word probe task	fMRI block design	Visual
	Pallier et al., 2011	Left TP, TPJ, anterior STS	Probe sentence, press a button on the basis of previous trial; word memory test at the end of each run	fMRI event-related design	Visual
	Shain et al., 2021	Left IFG, left MFG, left ATL, left PTL, left AG	No task?	fMRI event-related design	Visual

(Continued)

TABLE 2 (Continued)

Paradigm	References	Results	Task	Acquisition method	Stimuli presentation
Minimal phrases	Bemis and Pylkkänen, 2011 Adj-N vs. nonW-N	Increased activity in left ATL (84–225 ms) and vmPFC (300–500 ms)	Participants saw colored images in *composition* task and a colored blob and an outline in *non-composition* task: decide whether all words in the previous trial match the image	MEG block design	Visual
	Bemis and Pylkkänen, 2013a Adj-N vs. NonW-N	Left ATL (200–250 ms) regardless of word order in the compose task and for canonical word order in the non-compose task	Participants saw colored images in *composition* task and a colored blob and an outline in *non-composition* task: decide whether all words in the previous trial match the image	MEG between-subjects block design	Visual
	Bemis and Pylkkänen, 2013b Adj-N vs. nonW-N	Left ATL (191–299 ms visual modality; 268–323 ms auditory); left AG (336–390 ms in visual modality, 537–591 ms in auditory modality)	Participants saw colored images in *composition* task and a colored blob and an outline in *non-composition* task: decide whether all words in the previous trial match the image	MEG block design	Visual and auditory
	Fló et al., 2020 Experiment 1:Adj-N vs. nonW-noun	Negativities at 260–55 ms and 410–600 ms after word onset	Participants saw colored images after each trial; *composition* task: decide whether both words in previous trial match the image; *non-composition* task: decide if any of the preceding words match the image	EEG block design	Visual
	Fló et al., 2020 Experiment 2:Adj-N vs. nonW-noun	No effect of composition	Decide whether the image after each trial matches the preceding material	EEG event-related design	Visual
	Fritz and Baggio, 2020, 2022 Adj-N vs. nonW-N and pseudoW-N	450–700 ms positivity over centro-parietal channels (P600 ERP)	Comprehension questions after each trial	EEG event-related design	Visual
	Kochari et al., 2021 Adj-N vs. nonW-N	No effect	One or two words followed by a question mark; participants had to convert them into questions and answer	MEG event-related design	Visual
	Neufeld et al., 2016 Adj-N vs. nonW/pseudoW-N	Anterior negativity—50–100 ms starting at the first word; centro-parietal negativity after onset of second word (180–400 ms)	Participants saw colored images after each trial; *composition* task: decide whether both words in the previous trial match the image; *non-composition* task: decide if any of the preceding words matches the image	EEG block design	Visual
	Graessner et al., 2021b Adj-N vs. Adj-pseudoW	Task independent: left posterior AG, left posterior ITG; dorsomedial PFC. Explicit task: left anterior IFG, left ATL, left posterior MTG, left posterior AG, dorsomedial PFC, cerebellum. Implicit task: left AG, left posterior MTG/ITG, dorsomedial PFC	Session 1: implicit task: indicate whether both words had the same or different lexical status. Session 2: explicit task: indicate whether the phrase is meaningful or not	fMRI event-related design	Auditory

(Continued)

10.3389/flang.2023.1096110

TABLE 2 (Continued)

Paradigm	References	Results	Task	Acquisition method	Stimuli presentation
	Murphy et al., 2022 Adj-N vs. pseudoW-N vs. adj-pseudoW	Broadband gamma activity 210 ms after noun onset in portions of posterior STS	Participants saw colored pictures after each trial: decide whether the picture fully matches the previous phrase	iEEG/ECoG event-related design	Auditory
	Kim and Pylkkänen, 2019 Adverb-verb vs. nonW-verb	Increased activity at 250 ms in left ATL for eventive adverbs, in right ATL for agentive adverbs	Participants chose among two nouns which one fit best the meaning of the previous phrase	MEG event-related design	Visual
	Boylan et al., 2015 V-N/P-N/V-adv/adj-N vs. N/V-nonW	Activation in left and right AG for phrases sharing a verb regardless of composition type	Press a button indicating whether a two-word phrase was synonymous with the previous trial (30% of trials)	fMRI event-related design	Visual
	Zaccarella et al., 2017 PP-Det-N vs. 3-word list	Left IFG (BA44), left pSTS	Categorize the type of the previous trial (sentence, phrase word list, "rubbish")	fMRI block design	visual
	Matchin et al., 2019b V-Det-N vs. Det-A-N vs. lists of 3 words	Composition: left AG, left ATL, left posterior STS, left anterior IFG VP > NP: left posterior IFG, left posterior STS	Phrase probe task. After sequences of 3 trials participants saw a probe similar to a previous trial with one word changed; decide whether the probe is synonymous with one of the preceding trials	fMRI block design	Visual
	Westerlund et al., 2015 Modification: adj-N vs. nonW-N; adv-V vs. nonW-V; adv-adj vs. nonW-ADJ Argument saturation: V-N vs. nonW-N; P-N vs. nonW-N; det-N vs. nonW-N	Left ATL activation around 250 ms after second word onset for both composition types, but earlier for saturation	Phrase probe task after 20% of trials; indicate whether the probe is related to the previous trial	MEG event-related design	Visual
	Strijkers et al., 2019 PersPron-V/PossPron-N vs. ###-N/V	Activity in left and right IFG (starting 80 ms after second word) for N vs. V in combinatorial conditions only	Detect catch phrases (second word is a pseudoW)	MEG event-related design	Visual
	Kim and Pylkkänen, 2021 N-V-N vs. ###-V-N vs. ###-###-N	Subject-verb composition: left ATL (313–376 ms), left middle STC (332–364 ms); no effect for verb-object composition	Decide whether a picture presented after each trial accurately describes the linguistic material in the previous trial	MEG event-related design	Visual
	Lau and Liao, 2018 Adj-N and adj-N vs. adj-N ### adj-N	Increased anterior negativity starting at the first word of the second phrase lasting throughout the epoch	Memory probe of two words (20% of trials)	EEG block design (Experiment 1); event-related design (Experiment 2)	Visual
	Schell et al., 2017 A-N vs. N; Det-N vs. N	Adj-N composition: left IFG (BA45), left AG; det-N composition: left IFG (BA44), left posterior STS	Decide whether the previous trial could be integrated in a normal sentence	fMRI event-related design	Auditory

The paradigms included are those reviewed in Section 2. We report the results for the comparisons between well-formed meaningful sentences or phrases and the relevant baselines (specified in columns 1 or 2).

4. Weighing the options: What are we left with?

We have reviewed studies using different paradigms that tried to isolate composition in brain signals. The limitations of the paradigms discussed here are not entirely unknown and have been occasionally pointed out before (e.g., see Humphries et al., 2006; Matchin et al., 2017, 2019a). In this section, we reflect on what has been achieved so far in mapping semantic composition in brain space and time (for an earlier assessment, see Baggio, 2018). Table 1 summarizes our evaluation of the paradigms discussed above, and Table 2 is an overview of the main results of different studies. Our recommendation for the field includes developing new paradigms that overcome the limitations of current ones. A parallel strategy is to integrate results *across* studies and paradigms, in the hope that paradigms with complementary strengths and limitations would support each other and allow more reliable inferences from data. We briefly pursue this avenue here.

Despite their limitations, the words list and scrambled sentence paradigms allow lexical variables between stimuli to be matched. Although comparing sentences with scrambled versions may result in loss of signal (see above), scrambled sentences should still involve "less composition". Results from studies using this paradigm could help narrow down the search space of correlates of composition: regions engaged across studies using different baselines are candidate correlates of composition; regions that differ across studies may be related to processing of the particular stimuli used. The left posterior STS/STG, ATL, and AG consistently show up in normal vs. scrambled sentences contrasts. The left IFG is active in studies with difficult or engaging tasks, in studies using lists without function words, or words not in the original sentences. Further research is needed to understand how different baselines affect comparisons with normal sentences.

Jabberwocky sentences are a clever way of disentangling syntax from semantics, though formal aspects of meaning remain in stimuli with real function words and affixes. In this sense, like lists of words, Jabberwocky may involve semantic composition, but to a lower degree. Studies using this design often either reveal regions that overlap with those from studies using word lists or no effects in comparisons to sentences. Negative findings may suggest that lists are a better baseline than Jabberwocky, while overlapping results may indicate either that they are both equally effective or that both have issues with the same impact on brain signals. Minimal phrase designs using real word lists or pseudowords in baseline conditions have arguably made the most progress in narrowing down the space of correlates of composition. Zaccarella et al. (2017) and Matchin et al. (2017) implicate left IFG and pSTS in composition, while Murphy et al. (2022) localize effects of phrasal composition in pSTS around 200–300 ms from word onset. Inconsistencies remain across studies using minimal phrases as to the regions involved (left IFG, AG, vmPFC), with one frequently reported region being the left ATL. Yet, the LATL is mostly sensitive to *conceptual composition*. Integrating results from the studies in Table 2 we thus find a network in the left perisylvian cortex, with possibly the most functionally critical node in the posterior superior temporal gyrus and sulcus.

Section 3 considers alternative strategies, including testing theoretical distinctions Section 3.1, using advanced analysis methods Section 3.2, and developing new paradigms Section 3.3. We believe that initiating a discussion on the need to refine our paradigms is a crucial step forward, but a combination of approaches, as suggested in Section 3, as well as comparing results across methods (Table 2), is already leading to testable new hypotheses about the likely cortical seats and time course of syntax-driven meaning composition.

Our assessment of the different paradigms in Table 1 suggests that they are not all equal in their strengths and limitations. But the important lesson here is that while paradigms can be assessed on design grounds alone, they must also be evaluated *empirically* based on the plausibility and consistency of the results they generate: it is impossible to know exactly how the brain reacts to the different conditions a priori, and thus how severe the issues identified a priori may actually be. Comparing results across different paradigms can not only help us restrict the search of correlates of composition to fewer candidates: it can also provide indirect evidence of the actual impact of the limitations of particular paradigms. That said, this complex evaluative exercise remains fraught with difficulties, and is ultimately based on researcher choices, expertise, and judgement. For this reason, the way forward for the field should also involve the development of new paradigms and cannot be based entirely on comparison and integration of results across existing ones.

5. Conclusion

This review has examined experimental paradigms and designs used to search for neural correlates of syntax-driven meaning composition. Our aim was to dissect each paradigm presenting the ways in which it has been implemented in specific studies, bringing forth its goals and assumptions, and uncovering its strengths and weaknesses. One conclusion concerns the lack of baseline or control conditions that can fully prevent composition at specific points in time. Without such conditions, interpreting comparisons with phrases or sentences remains difficult: any claim that a given signal is a correlate of composition is undermined, if the conditions compared do not *only* differ in whether composition is engaged or not. This may partly explain why M/EEG or fMRI studies have not revealed correlates of composition invariant across studies or paradigms (Table 2). But as noted, the challenge ultimately involves more than just experimental design: finding the neural mechanisms of composition will also require progress in integrative theory (van Rooij and Baggio, 2020, 2021), recording resolution, and data acquisition and analysis.

Here, we have focused on a neglected, yet essential ingredient of research methodology: the internal validity of experimental paradigms and designs. Our critique is not meant to devalue the ingenuity of experimental designs used by researchers throughout the years: we have contributed to this research ourselves, and we have used several of the classical paradigms in our work. Some of the issues raised here were also noted by others, but we believe it is useful to assess different paradigms comparatively and systematically, using the same standards. In addition to examining the limitations of baseline conditions, we should

reconsider the theoretical assumptions about composition that we build into our experimental designs. The brain may not always automatically compute meaning taking syntactic structure into account: if syntax is not always deployed during comprehension, or if lexical processing in well-formed and meaningful sentences always engages a set of independent operations in addition to syntax-driven composition, then any comparison of conditions, even assuming adequate baselines, will reveal either less or more in terms of neural signals than syntax-driven composition (Baggio, 2018, 2022).

Consider the "standard view" of meaning composition from generative syntax and formal semantics. As a computational implementation of composition, that view may not quite provide what psycholinguists and neurolinguists need to derive specific predictions and explain existing experimental results. One reason is that there is still no real consensus on the atoms and structures of syntax in the first place, their relation to lexical encoding, and the semantic primitives of combination they correspond to. The logical calculus of formal semantics works equally well for very different choices of syntactic and semantic ontology: the existence of a syntax-semantics interface that respects function-argument composition does not, in and of itself, provide a unique answer to what those syntactic and semantic primitives are. Moreover, there are indications that the basic combinatoric building blocks assumed in formal semantic theory do not map in any systematic way to basic differences and measures at the neurolinguistic level (Pylkkänen and McElree, 2006).

Putting aside open questions of what the minimal parts and modes of combination are, one could disagree with the particulars of this narrow formulation, and specifically with the centrality of Compositionality (Baggio et al., 2012b; Baggio, 2018, 2021). But the key insight here is that human languages have algorithms for building meanings predictably from their parts. Predictability and generativity of meaning should be taken seriously as computational constraints modulating language processing and its outputs, even though not all complex meanings may be equally subject to Compositionality. Developing better experimental paradigms should go hand in hand with theoretical and modeling efforts aimed at charting the different ways in which brains actually build meaning.

Author contributions

LC did the literature search and selected the relevant studies. LC, GB, and GR wrote and edited the paper. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adger, D. (2003). *Core Syntax: A Minimalist Approach*. Oxford: Oxford University Press.

Artoni, F., d'Orio, P., Catricalà, E., Conca, F., Bottoni, F., Pelliccia, V., et al. (2020). High gamma response tracks different syntactic structures in homophonous phrases. *Sci. Rep.* 10, 1–10. doi: 10.1038/s41598-020-64375-9

Baggio, G. (2012). Selective alignment of brain responses by task demands during semantic processing. *Neuropsychologia* 50, 655–665. doi: 10.1016/j.neuropsychologia.2012.01.002

Baggio, G. (2018). *Meaning in the Brain*. Cambridge MA: MIT Press.

Baggio, G. (2020). "Epistemic transfer between linguistics and neuroscience: problems and prospects," in *The Philosophy and Science of Language* (Palgrave Macmillan), 275–308.

Baggio, G. (2021). Compositionality in a parallel architecture for language processing. *Cogn. Sci.* 45, e12949. doi: 10.1111/cogs.12949

Baggio, G. (2022). *Neurolinguistics*. Cambridge MA: MIT Press.

Baggio, G., Choma, T., van Lambalgen, M., and Hagoort, P. (2010). Coercion and compositionality. *J. Cogn. Neurosci.* 22, 2131–2140. doi: 10.1162/jocn.2009.21303

Baggio, G., and Hagoort, P. (2011). The balance between memory and unification in semantics: a dynamic account of the N400. *Lang. Cogn. Process.* 26, 1338–1367. doi: 10.1080/01690965.2010.542671

Baggio, G., Stenning, K., and van Lambalgen, M. (2016). "Semantics and cognition," in *The Cambridge Handbook of Formal Semantics*, eds M. Aloni, and P. Dekker (Cambridge: Cambridge University Press), 756–774.

Baggio, G., van Lambalgen, M., and Hagoort, P. (2008). Computing and recomputing discourse models: an ERP study. *J. Mem. Lang.* 59, 36–53. doi: 10.1016/j.jml.2008.02.005

Baggio, G., van Lambalgen, M., and Hagoort, P. (2012a). "Language, linguistics and cognition," in *Philosophy of Linguistics*, eds R. Kempson, T. Fernando, and N. Asher (Amsterdam: Elsevier), 325–355.

Baggio, G., van Lambalgen, M., and Hagoort, P. (2012b). "The processing consequences of compositionality," in *The Oxford Handbook of Compositionality*, eds M. Werning, W. Hinzen, and E. Machery (Oxford: Oxford University Press), 655–672.

Bautista, A., and Wilson, S. M. (2016). Neural responses to grammatically and lexically degraded speech. *Lang. Cognit. Neurosci.* 31, 567–574. doi: 10.1080/23273798.2015.1123281

Bemis, D. K., and Pylkkänen, L. (2011). Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *J. Neurosci.* 31, 2801–2814. doi: 10.1523/JNEUROSCI.5003-10.2011

Bemis, D. K., and Pylkkänen, L. (2013a). Flexible composition: MEG evidence for the deployment of basic combinatorial linguistic mechanisms in response to task demands. *PLoS ONE* 8, e73949. doi: 10.1371/journal.pone.0073949

Bemis, D. K., and Pylkkänen, L. (2013b). Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex* 23, 1859–1873. doi: 10.1093/cercor/bhs170

Boylan, C., Trueswell, J. C., and Thompson-Schill, S. L. (2015). Compositionality and the angular gyrus: a multi-voxel similarity analysis of the semantic composition of nouns and verbs. *Neuropsychologia* 78, 130–141. doi: 10.1016/j.neuropsychologia.2015.10.007

Boylan, C., Trueswell, J. C., and Thompson-Schill, S. L. (2017). Relational vs. attributive interpretation of nominal compounds differentially engages angular gyrus and anterior temporal lobe. *Brain Lang.* 169, 8–21. doi: 10.1016/j.bandl.2017.01.008

Branco, P., Seixas, D., and Castro, S. L. (2020). Mapping language with resting-state functional magnetic resonance imaging: a study on the functional profile of the language network. *Hum. Brain Mapp.* 41, 545–560. doi: 10.1002/hbm.24821

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., and Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain Lang.* 120, 163–173. doi: 10.1016/j.bandl.2010.04.002

Brennan, J., and Pylkkänen, L. (2012). The time-course and spatial distribution of brain activity associated with sentence processing. *Neuroimage* 60, 1139–1148. doi: 10.1016/j.neuroimage.2012.01.030

Brennan, J. R., and Pylkkänen, L. (2017). MEG evidence for incremental sentence composition in the anterior temporal lobe. *Cogn. Sci.* 41, 1515–1531. doi: 10.1111/cogs.12445

Burroughs, A., Kazanina, N., and Houghton, C. (2021). Grammatical category and the neural processing of phrases. *Sci. Rep.* 11, 2446. doi: 10.1038/s41598-021-81901-5

Coopmans, C. W., de Hoop, H., Hagoort, P., and Martin, A. E. (2022). Effects of structure and meaning on cortical tracking of linguistic units in naturalistic speech. *Neurobiol. Lang.* 3, 386–412. doi: 10.1162/nol_a_00070

Culicover, P. W., and Jackendoff, R. (2006). The simpler syntax hypothesis. *Trends Cogn. Sci.* 10, 413–418. doi: 10.1016/j.tics.2006.07.007

Del Prato, P., and Pylkkänen, L. (2014). MEG evidence for conceptual combination but not numeral quantification in the left anterior temporal lobe during language production. *Front. Psychol.* 5, 524. doi: 10.3389/fpsyg.2014.00524

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nat. Neurosci.* 19, 158–164. doi: 10.1038/nn.4186

Fedorenko, E., Blank, I. A., Siegelman, M., and Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition* 203, 104348. doi: 10.1016/j.cognition.2020.104348

Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S., and Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol.* 104, 1177–1194. doi: 10.1152/jn.00032.2010

Fedorenko, E., Nieto-Castanon, A., and Kanwisher, N. (2012). Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia* 50, 499–513. doi: 10.1016/j.neuropsychologia.2011.09.014

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., et al. (2016). Neural correlate of the construction of sentence meaning. *Proc. Nat. Acad. Sci. U. S. A.* 113, E6256–E6262. doi: 10.1073/pnas.1612132113

Flick, G., and Pylkkänen, L. (2020). Isolating syntax in natural language: MEG evidence for an early contribution of left posterior temporal cortex. *Cortex* 127, 42–57. doi: 10.1016/j.cortex.2020.01.025

Fló, E., Cabana, Á., and Valle-Lisboa, J. C. (2020). EEG signatures of elementary composition: disentangling genuine composition and expectancy processes. *Brain Lang.* 209, 104837. doi: 10.1016/j.bandl.2020.104837

Friederici, A. D., Gunter, T. C., Hahne, A., and Mauth, K. (2004). The relative timing of syntactic and semantic processes in sentence comprehension. *Neuroreport* 15, 165–169. doi: 10.1097/00001756-200401190-00032

Friederici, A. D., Meyer, M., and Von Cramon, D. Y. (2000). Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information. *Brain Lang.* 74, 289–300. doi: 10.1006/brln.2000.2313

Fritz, I., and Baggio, G. (2020). Meaning composition in minimal phrasal contexts: distinct ERP effects of intensionality and denotation. *Lang. Cognit. Neurosci.* 35, 1295–1313. doi: 10.1080/23273798.2020.1749678

Fritz, I., and Baggio, G. (2022). Neural and behavioural effects of typicality, denotation and composition in an adjective–noun combination task. *Lang. Cognit. Neurosci.* 37, 537–559. doi: 10.1080/23273798.2021.2004176

Glushko, A., Poeppel, D., and Steinhauer, K. (2022). Overt and implicit prosody contribute to neurophysiological responses previously attributed to grammatical processing. *Sci. Rep.* 12, 14759. doi: 10.1038/s41598-022-18162-3

Goucha, T., and Friederici, A. D. (2015). The language skeleton after dissecting meaning: a functional segregation within Broca's area. *Neuroimage* 114, 294–302. doi: 10.1016/j.neuroimage.2015.04.011

Graessner, A., Zaccarella, E., Friederici, A. D., Obrig, H., and Hartwigsen, G. (2021a). Dissociable contributions of frontal and temporal brain regions to basic semantic composition. *Brain Commun.* 3, fcab090. doi: 10.1093/braincomms/fcab090

Graessner, A., Zaccarella, E., and Hartwigsen, G. (2021b). Differential contributions of left-hemispheric language regions to basic semantic composition. *Brain Struct. Funct.* 226, 501–518. doi: 10.1007/s00429-020-02196-2

Grodzinsky, Y., Pieperhoff, P., and Thompson, C. (2021). Stable brain loci for the processing of complex syntax: A review of the current neuroimaging evidence. *Cortex* 142, 252–271. doi: 10.1016/j.cortex.2021.06.003

Hagoort, P., Baggio, G., and Willems, R. M. (2009). "Semantic unification," in *The Cognitive Neurosciences, 4th Edn*, ed M. Gazzaniga (Cambridge MA: MIT Press), 819–836.

Hagoort, P., and Indefrey, P. (2014). The neurobiology of language beyond single words. *Annu. Rev. Neurosci.* 37, 347–362. doi: 10.1146/annurev-neuro-071013-013847

Hardy, S. M., Jensen, O., Wheeldon, L., Mazaheri, A., and Segaert, K. (2023). Modulation in alpha band activity reflects syntax composition: an MEG study of minimal syntactic binding. *Cerebral Cortex* 33, 497–511. doi: 10.1093/cercor/bhac080

Hashimoto, R., and Sakai, K. L. (2002). Specialization in the left prefrontal cortex for sentence comprehension. *Neuron* 35, 589–597. doi: 10.1016/S0896-6273(02)00788-2

Hasson, U., Egidi, G., Marelli, M., and Willems, R. M. (2018). Grounding the neurobiology of language in first principles: the necessity of non-language-centric explanations for language comprehension. *Cognition* 180, 135–157. doi: 10.1016/j.cognition.2018.06.018

Heim, I., and Kratzer, A. (1998). *Semantics in Generative Grammar*. Oxford: Blackwell.

Heim, S., Alter, K., Ischebeck, A. K., Amunts, K., Eickhoff, S. B., Mohlberg, H., et al. (2005). The role of the left Brodmann's areas 44 and 45 in reading words and pseudowords. *Cogn. Brain Res.* 25, 982–993. doi: 10.1016/j.cogbrainres.2005.09.022

Hultén, A., Schoffelen, J. M., Uddén, J., Lam, N. H., and Hagoort, P. (2019). How the brain makes sense beyond the processing of single words–An MEG study. *Neuroimage* 186, 586–594. doi: 10.1016/j.neuroimage.2018.11.035

Humphries, C., Binder, J. R., Medler, D. A., and Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *J. Cogn. Neurosci.* 18, 665–679. doi: 10.1162/jocn.2006.18.4.665

Humphries, C., Love, T., Swinney, D., and Hickok, G. (2005). Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Hum. Brain Mapp.* 26, 128–138. doi: 10.1002/hbm.20148

Husband, E. M., Kelly, L. A., and Zhu, D. C. (2011). Using complement coercion to understand the neural basis of semantic composition: evidence from an fMRI study. *J. Cogn. Neurosci.* 23, 3254–3266. doi: 10.1162/jocn_a_00040

Iwabuchi, T., and Makuuchi, M. (2021). When a sentence loses semantics: Selective involvement of a left anterior temporal subregion in semantic processing. *Eur. J. Neurosci.* 53, 929–942. doi: 10.1111/ejn.15022

Kalenkovich, E., Shestakova, A., and Kazanina, N. (2022). Frequency tagging of syntactic structure or lexical properties; a registered MEG study. *Cortex* 146, 24–38. doi: 10.1016/j.cortex.2021.09.012

Kaufeld, G., Bosker, H. R., Ten Oever, S., Alday, P. M., Meyer, A. S., and Martin, A. E. (2020). Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *J. Neurosci.* 40, 9467–9475. doi: 10.1523/JNEUROSCI.0302-20.2020

Kazanina, N., and Tavano, A. (2023). What neural oscillations can and cannot do for syntactic structure building. *Nat. Rev. Neurosci.* 24, 113–128. doi: 10.1038/s41583-022-00659-5

Kim, S., and Pylkkänen, L. (2019). Composition of event concepts: evidence for distinct roles for the left and right anterior temporal lobes. *Brain Lang.* 188, 18–27. doi: 10.1016/j.bandl.2018.11.003

Kim, S., and Pylkkänen, L. (2021). How the conceptual specificity of individual words affects incremental sentence composition: MEG evidence. *Brain Lang.* 218, 104951. doi: 10.1016/j.bandl.2021.104951

Kochari, A. R., Lewis, A. G., Schoffelen, J. M., and Schriefers, H. (2021). Semantic and syntactic composition of minimal adjective-noun phrases in Dutch: an MEG study. *Neuropsychologia* 155, 107754. doi: 10.1016/j.neuropsychologia.2021.107754

Kuperberg, G. R., Choi, A., Cohn, N., Paczynski, M., and Jackendoff, R. (2010). Electrophysiological correlates of complement coercion. *J. Cogn. Neurosci.* 22, 2685–2701. doi: 10.1162/jocn.2009.21333

Kuperberg, G. R., McGuire, P. K., Bullmore, E. T., Brammer, M. J., Rabe-Hesketh, S., Wright, I. C., et al. (2000). Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: an fMRI study. *J. Cogn. Neurosci.* 12, 321–341. doi: 10.1162/089892900562138

Lau, E., and Liao, C. H. (2018). Linguistic structure across time: ERP responses to coordinated and uncoordinated noun phrases. *Lang. Cogn. Neurosci.* 33, 633–647. doi: 10.1080/23273798.2017.1400081

Law, R., and Pylkkänen, L. (2021). Lists with and without syntax: a new approach to measuring the neural processing of syntax. *J. Neurosci.* 41, 2186–2196. doi: 10.1523/JNEUROSCI.1179-20.2021

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* 31, 2906–2915. doi: 10.1523/JNEUROSCI.3684-10.2011

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman.

Matar, S., Dirani, J., Marantz, A., and Pylkkänen, L. (2021). Left posterior temporal cortex is sensitive to syntax within conceptually matched Arabic expressions. *Sci. Rep.* 11, 1–14. doi: 10.1038/s41598-021-86474-x

Matchin, W., Brodbeck, C., Hammerly, C., and Lau, E. (2019a). The temporal dynamics of structure and content in sentence comprehension: evidence from fMRI-constrained MEG. *Hum. Brain Mapp.* 40, 663–678. doi: 10.1002/hbm.24403

Matchin, W., Hammerly, C., and Lau, E. (2017). The role of the IFG and pSTS in syntactic prediction: evidence from a parametric study of hierarchical structure in fMRI. *Cortex* 88, 106–123. doi: 10.1016/j.cortex.2016.12.010

Matchin, W., Liao, C. H., Gaston, P., and Lau, E. (2019b). Same words, different structures: an fMRI investigation of argument relations and the angular gyrus. *Neuropsychologia* 125, 116–128. doi: 10.1016/j.neuropsychologia.2019.01.019

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., et al. (1993). The cortical representation of speech. *J. Cogn. Neurosci.* 5, 467–479. doi: 10.1162/jocn.1993.5.4.467

Meyer, L., Sun, Y., and Martin, A. E. (2020). Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Lang. Cogn. Neurosci.* 35, 1089–1099. doi: 10.1080/23273798.2019.1693050

Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., et al. (2020). Composition is the core driver of the language-selective network. *Neurobiol. Lang.* 1, 104–134. doi: 10.1162/nol_a_00005

Murphy, E. (2020). *The Oscillatory Nature of Language*. Cambridge: Cambridge University Press.

Murphy, E., Woolnough, O., Rollo, P. S., Roccaforte, Z. J., Segaert, K., Hagoort, P., et al. (2022). Minimal phrase composition revealed by intracranial recordings. *J. Neurosci.* 42, 3216–3227. doi: 10.1523/JNEUROSCI.1575-21.2022

Nefdt, R. M., and Baggio, G. (2023). Notational variants and cognition: the case of dependency grammar. *Erkenntnis* 1–31. doi: 10.1007/s10670-022-00657-0

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., et al. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proc. Nat. Acad. Sci. U. S. A.* 114, E3669–E3678. doi: 10.1073/pnas.1701590114

Neufeld, C., Kramer, S. E., Lapinskaya, N., Heffner, C. C., Malko, A., and Lau, E. F. (2016). The electrophysiology of basic phrase building. *PLoS ONE* 11, e0158446. doi: 10.1371/journal.pone.0158446

Ni, W., Constable, R. T., Mencl, W. E., Pugh, K. R., Fulbright, R. K., Shaywitz, S., et al. (2000). An event-related neuroimaging study distinguishing form and content in sentence processing. *J. Cogn. Neurosci.* 12, 120–133. doi: 10.1162/08989290051137648

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., et al. (2020). Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. *Philos. Transact. R. Soc. B* 375, 20180522. doi: 10.1098/rstb.2018.0522

Olstad, A. M. H., Fritz, I., and Baggio, G. (2020). Composition decomposed: distinct neural mechanisms support processing of nouns in modification and predication contexts. *J. Exp. Psychol. Learn. Mem. Cognit.* 46, 2193. doi: 10.1037/xlm0000943

Paczynski, M., Jackendoff, R., and Kuperberg, G. (2014). When events change their nature: the neurocognitive mechanisms underlying aspectual coercion. *J. Cogn. Neurosci.* 26, 1905–1917. doi: 10.1162/jocn_a_00638

Pallier, C., Devauchelle, A. D., and Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proc. Nat. Acad. Sci. U. S. A.* 108, 2522–2527. doi: 10.1073/pnas.1018711108

Parrish, A., and Pylkkänen, L. (2022). Conceptual combination in the LATL with and without syntactic composition. *Neurobiol. Lang.* 3, 46–66. doi: 10.1162/nol_a_00048

Piñango, M. M., and Deo, A. (2016). Reanalyzing the complement coercion effect through a generalized lexical semantics for aspectual verbs. *J. Semant.* 33, 359–408. doi: 10.1093/jos/ffv003

Pylkkänen, L., and McElree, B. (2006). "The syntax-semantics interface: on-line composition of sentence meaning," in *Handbook of Psycholinguistics*, eds M. Traxler and M. Gernsbacher (Amsterdam: Elsevier/Academic Press), 539–579.

Pylkkänen, L. (2008). Mismatching meanings in brain and behavior. *Lang. Linguist. Compass* 2, 712–738. doi: 10.1111/j.1749-818X.2008.00073.x

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science* 366, 62–66. doi: 10.1126/science.aax0050

Pylkkänen, L., Brennan, J., and Bemis, D. K. (2011). Grounding the cognitive neuroscience of semantics in linguistic theory. *Lang. Cogn. Process.* 26, 1317–1337. doi: 10.1080/01690965.2010.527490

Pylkkänen, L., and Brennan, J. R. (2019). Composition: the neurobiology of syntactic and semantic structure building. *PsyArXiv. [Preprint].* doi: 10.31234/osf.io/fa2xb

Pylkkänen, L., and McElree, B. (2007). An MEG study of silent meaning. *J. Cogn. Neurosci.* 19, 1905–1921. doi: 10.1162/jocn.2007.19.11.1905

Röder, B., Stock, O., Neville, H., Bien, S., and Rösler, F. (2002). Brain activation modulated by the comprehension of normal and pseudo-word sentences of different processing demands: a functional magnetic resonance imaging study. *Neuroimage* 15, 1003–1014. doi: 10.1006/nimg.2001.1026

Schell, M., Zaccarella, E., and Friederici, A. D. (2017). Differential cortical contribution of syntax and semantics: an fMRI study on two-word phrasal processing. *Cortex* 96, 105–120. doi: 10.1016/j.cortex.2017.09.002

Schumacher, P. B. (2013). When combinatorial processing results in reconceptualization: toward a new approach of compositionality. *Front. Psychol.* 4, 677. doi: 10.3389/fpsyg.2013.00677

Segaert, K., Mazaheri, A., and Hagoort, P. (2018). Binding language: structuring sentences through precisely timed oscillatory mechanisms. *Eur. J. Neurosci.* 48, 2651–2662. doi: 10.1111/ejn.13816

Shain, C., Kean, H., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., et al. (2021). 'Constituent length' effects in fMRI do not provide evidence for abstract syntactic processing. *BioRxiv [Preprint]*. doi: 10.1101/2021.11.12.467812

Sheng, J., Zheng, L., Lyu, B., Cen, Z., Qin, L., Tan, L. H., et al. (2019). The cortical maps of hierarchical linguistic structures during speech perception. *Cerebral Cortex* 29, 3232–3240. doi: 10.1093/cercor/bhy191

Strijkers, K., Chanoine, V., Munding, D., Dubarry, A. S., Trébuchon, A., Badier, J. M., et al. (2019). Grammatical class modulates the (left) inferior frontal gyrus within 100 milliseconds when syntactic context is predictive. *Sci. Rep.* 9, 1–13. doi: 10.1038/s41598-019-41376-x

Stromswold, K., Caplan, D., Alpert, N., and Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain Lang.* 52, 452–473. doi: 10.1006/brln.1996.0024

van Rooij, I., and Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychol. Inq.* 31, 321–325. doi: 10.1080/1047840X.2020.1853477

van Rooij, I., and Baggio, G. (2021). Theory before the test: how to build high-verisimilitude explanatory theories in psychological science. *Perspect. Psychol. Sci.* 16, 682–697. doi: 10.1177/1745691620970604

Vandenberghe, R., Nobre, A. C., and Price, C. J. (2002). The response of left temporal cortex to sentences. *J. Cogn. Neurosci.* 14, 550–560. doi: 10.1162/08989290260045800

Westerlund, M., Kastner, I., Al Kaabi, M., and Pylkkänen, L. (2015). The LATL as locus of composition: MEG evidence from English and Arabic. *Brain Lang.* 141, 124–134. doi: 10.1016/j.bandl.2014.12.003

Williams, A., Reddigari, S., and Pylkkänen, L. (2017). Early sensitivity of left perisylvian cortex to relationality in nouns and verbs. *Neuropsychologia* 100, 131–143. doi: 10.1016/j.neuropsychologia.2017.04.029

Xu, J., Kemeny, S., Park, G., Frattali, C., and Braun, A. (2005). Language in context: emergent features of word, sentence, and narrative comprehension. *Neuroimage* 25, 1002–1015. doi: 10.1016/j.neuroimage.2004.12.013

Zaccarella, E., Meyer, L., Makuuchi, M., and Friederici, A. D. (2017). Building by syntax: the neural basis of minimal linguistic structures. *Cerebral Cortex* 27, 411–421. doi: 10.1093/cercor/bhv234

Zhang, L., and Pylkkänen, L. (2015). The interplay of composition and concept specificity in the left anterior temporal lobe: an MEG study. *Neuroimage* 111, 228–240. doi: 10.1016/j.neuroimage.2015.02.028

Ziegler, J., and Pylkkänen, L. (2016). Scalar adjectives and the temporal unfolding of semantic composition: an MEG investigation. *Neuropsychologia* 89, 161–171. doi: 10.1016/j.neuropsychologia.2016.06.010