Evidence for a compressed neural code in working memory during language composition

Or

Evidence of a compressed neural code of language composition during working memory

Or

Language composition: evidence of a compressed neural code in working memory

Abstract

The human brain is uniquely equipped to meaningfully combine successive elements on the spot. However, the representational format and the neural bases of such compositionality remains largely unknown. Here, we tackle this issue using magnetoencephalography recordings of brain activity while subjects compared 1-, 2- and 5-word phrases to a subsequent image. The decoding of MEG signals reveals three main findings: First, the representation of each word is robustly maintained until it is integrated with its corresponding phrase, and then fades away. Second, the neural activity during the delay period correlates with the complexity of the phrase. Third, the neural and behavioral read-out of the sentence depends on this complexity and is faster for surface properties of the phrase compared to syntactically deeper ones. Together, these results suggest that compositional representations are compressed in working memory and require task-specific decompression to be accessed. Overall, these results shed new light on the nature of compositional representations in the human brain.

Introduction

The ability to compose individual elements into a meaningful representation is arguably the paramount skill of the human mind, to the point that it has been called the “holy grail” of cognitive science (Jackendoff, 2002). It is formidable, though often overlooked that we are able to instantly understand sentences that we have never heard before, by effortlessly binding words in real time and infer their combined meaning.

There is however, no consensus regarding how the brain combines the meaning of individual elements, and how it represents such composition (Friederici et al., 2017; Martin & Doumas, 2017; Frankland & Greene, 2020). Among these theories, vector-symbolic architecture such as the tensor-product representation (Smolensky, 1990) and the semantic pointer architecture (Eliasmith & Anderson, 2003; Eliasmith et al., 2012) have seen notable success. For example, the semantic pointer architecture has been put to use in theories of concepts (Blouw et al., 2016), emotions (Kajić et al., 2019) and consciousness (Thagard &

Stewart, 2014). It was also found that representations of artificial neural networks could be well approximated by tensor-product representations when they are trained on artificial, explicitly compositional sequence-to-sequence tasks, but not when they are trained on natural language (McCoy et al., 2019; Soulos et al., 2020). Relatedly, recent successes in deep learning models of natural language processing seem to be partly due to their good generalization properties, although they do not rely on systematic compositional rules (Baroni, 2020; Brown et al., 2020; Chaabouni et al., 2020). For example, even state-of-the-art image generation models from natural language, such as OpenAI's Dall-E 2 (Radford et al., 2021; Ramesh et al., 2022) and Google's Imagen (Saharia et al., 2022) can dramatically fails on simple compositional and binding operations (Conwell & Ullman, 2022; Marcus et al., 2022)

Multiple brains regions have been associated with compositional processes. Famously, "Broca's area", more precisely the pars opercularis and triangularis of the inferior frontal gyrus (IFG) has long been thought to be the siege of unification operations (Hagoort, 2005), including composition and binding. This region was repeatedly found to be more active in conditions where composition could happened, such as sentences versus word lists (Mazoyer et al., 1993; Humphries et al., 2005; Friederici et al., 2010), normal sentences versus Jabberwocky (Fedorenko et al., 2016), and constituents of increasing size (Pallier et al., 2011; Nelson et al., 2017). In many of these studies, the posterior superior temporal sulcus (pSTS) was also found to be active.

In addition, the anterior temporal lobe (ATL) has been implicated over and over in two-word conceptual combinations (Bemis & Pylkkänen, 2011; Pylkkänen, 2019, 2020). ATL was also found to be more active when reading or listening to sentences compared to word list (J. Brennan & Pylkkänen, 2012), and to correlate with the operations of a syntactic parser in natural story reading paradigm (J. R. Brennan & Pylkkänen, 2017).

Recent studies on two-word compositions have found that the neural representation of the adjective is still present when its associated noun is presented, although this representation differs from its sensory representation (Fyshe et al., 2019; Honari-Jahromi et al., 2021). Whether a similar process happens for longer, more complex phrases remains unknown.

The idea that sequences are stored in an abstract compressed format started with foundational studies from Restle, who showed that people naturally decompose and store regular patterns as combinations of elementary rules (Restle, 1970; Restle & Brown, 1970). Much more recently, it has been shown that humans compress spatial sequences using geometrical primitives that can be decoded from MEG signals (Al Roumi et al., 2021). Similar compression operations were found in binary auditory and visual sequences (Planton et al., 2021). This framework has been suggested to apply in the context of natural language processing (Christiansen & Chater, 2016) but has no direct neural evidence for. Critically, the identification of such compressed representations in the brain activity, remains, to date, elusive.

This view is somewhat opposite to classical working memory theories where each characteristic of a stimulus is represented explicitly by sustained firing of specific neurons

(Goldman-Rakic, 1995; Leung et al., 2002). This same goes for more recent activity-silent theories of working memory, which posit that change in network-level characteristics (such as short-term plasticity) are the basis for the storage of short-term memoranda (Mongillo et al., 2008; Stokes, 2015), but no constraint is set on the representational format of the storage. Even recent proposal that combine sustained and activity-silent working memory (Spaak et al., 2017; Trübutschek et al., 2019; Barbosa et al., 2020; Stokes et al., 2020) consider that information is stored as is, not compressed, as suggested by Restle and others.

Relatedly, it has recently been proposed that intelligent behavior relies on factorized representations, where each property of the input is explicitly represented independently of the others (Behrens et al., 2018; Whittington et al., 2020). Such proposal is incompatible with the compression hypothesis: It predicts that accessing any characteristic of the memorandum would take approximately the same time.

Thus, we raise the questions: how long are individual words actively maintained in neural activity when subjects process and keep in mind a sentence? What is the format of the stored representations?

In this study we tackle the question of semantic composition using three tasks: a hybrid 1-back and two delayed sentence-to-image matching tasks (Figure 1), while magnetoencephalography (MEG) is recorded in 30 native French subjects. In one-feature blocks, the subjects underwent a 1-back word-image localizer in which the semantics of individual words had to be accessed, but no composition could occur (Figure 1A). In two- and five-features blocks, subjects read sentences describing objects composed of a shape (a noun, either “square”, “circle”, and “triangle”) and a color (an adjective that can be “blue”, “red”, or “green”) in a rapid visual serial presentation. In two-features blocks, a single object is presented (Figure 1B), whereas in five-features blocks, two objects are linked by a spatial relation (“to the left” or “to the right”, Figure 1C). After a delay, subjects were presented with an image and tasked to match it to the preceding sentence.

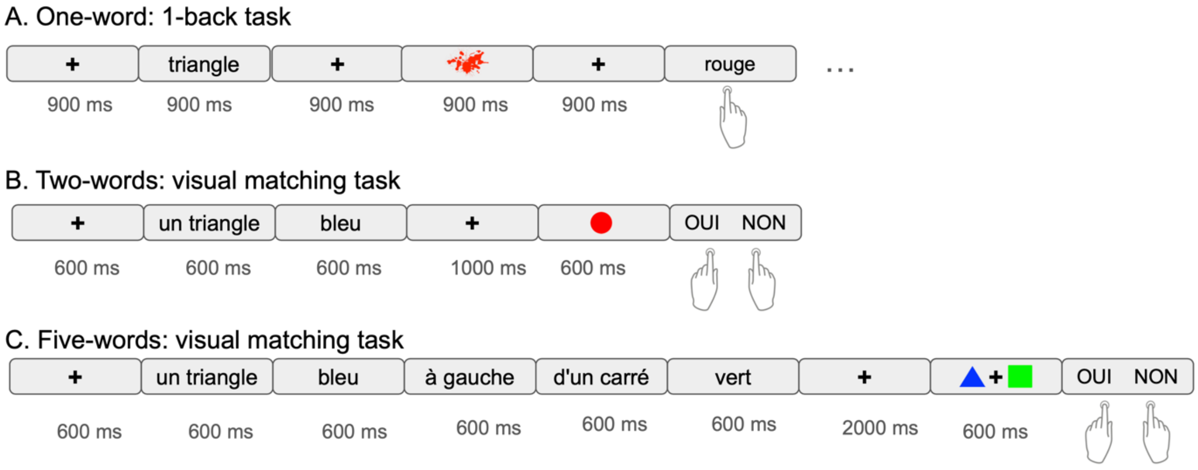


Figure 1: Experimental designs: hybrid 1-back and delayed sentence-to-image matching tasks
A: The one-word blocks consist of a 1-back across words and images. Subjects have to indicate, in a long series of random stimuli, when two stimuli represent the same meaning.
B: In two-words blocks, subjects have to determine whether the object described by the two

successive words match the image presented 1 s later.
C: Same as B but for 5 words sequences.

We use multivariate decoding and temporal generalization (King & Dehaene, 2014) to decipher the dynamics of composition. In short, decoding consists in learning a linear combination of activity from multiple sensors to try to predict experimental conditions, thus instructing us on whether the brain represents the condition of interest at the time of interest. Temporal generalization then assesses the ability of these classifiers to generalize to other time points than the one they were trained on, thus assessing whether the neural representations of experimental variables are stable over time. In other words, we use this within time decoding approach to study when the objects' shapes and colors are represented in the brain, and temporal generalization decoding to assess the stability of these representations.

We consider three pairs of hypotheses:

- Hypothesis 1 : if an active representation of each element is necessary for online composition, then individual words should be explicitly represented until they are combined with their corresponding phrase. On the other hand, if composition can occur solely with activity-silent mechanisms, then the active representation of each word until the end of its constituent would not be necessary.
- Hypothesis 1: if neural representations are compressed for short-term storage, then some neural signal should reflect the complexity (i.e., the quantity of information) of the sentence. To the contrary, if the representations are not compressed, such that redundant properties are encoded independently, then neural signals should not vary with complexity.
- Hypothesis 1: to decode such compressed representation, an active decompression operation should take place, with variable delays depending on the complexity and syntactic depth of the property that is being read-out. Otherwise, if neural representations are factorized then accessing each property of the memorandum should take approximately the same time.

Consequently, we focus on three phases: (1) stimulus presentation to look for online composition, (2) delay period, to study how these sentential representations are stored in working memory, and (3) image probe to examine how the representations are read-out.

Results

Words are maintained longer when they need to be combined with subsequent words

We start by examining the immediate dynamics of semantic composition, i.e., the activity during the presentation of the sentence. We trained logistic regressions to classify which i) shape, and ii) color was presented to the subjects in individual trials from each block type. The decoding performance rose around 200 ms after word onset and reached at least 0.55 AUC in all conditions, for each block type (Figure 2).

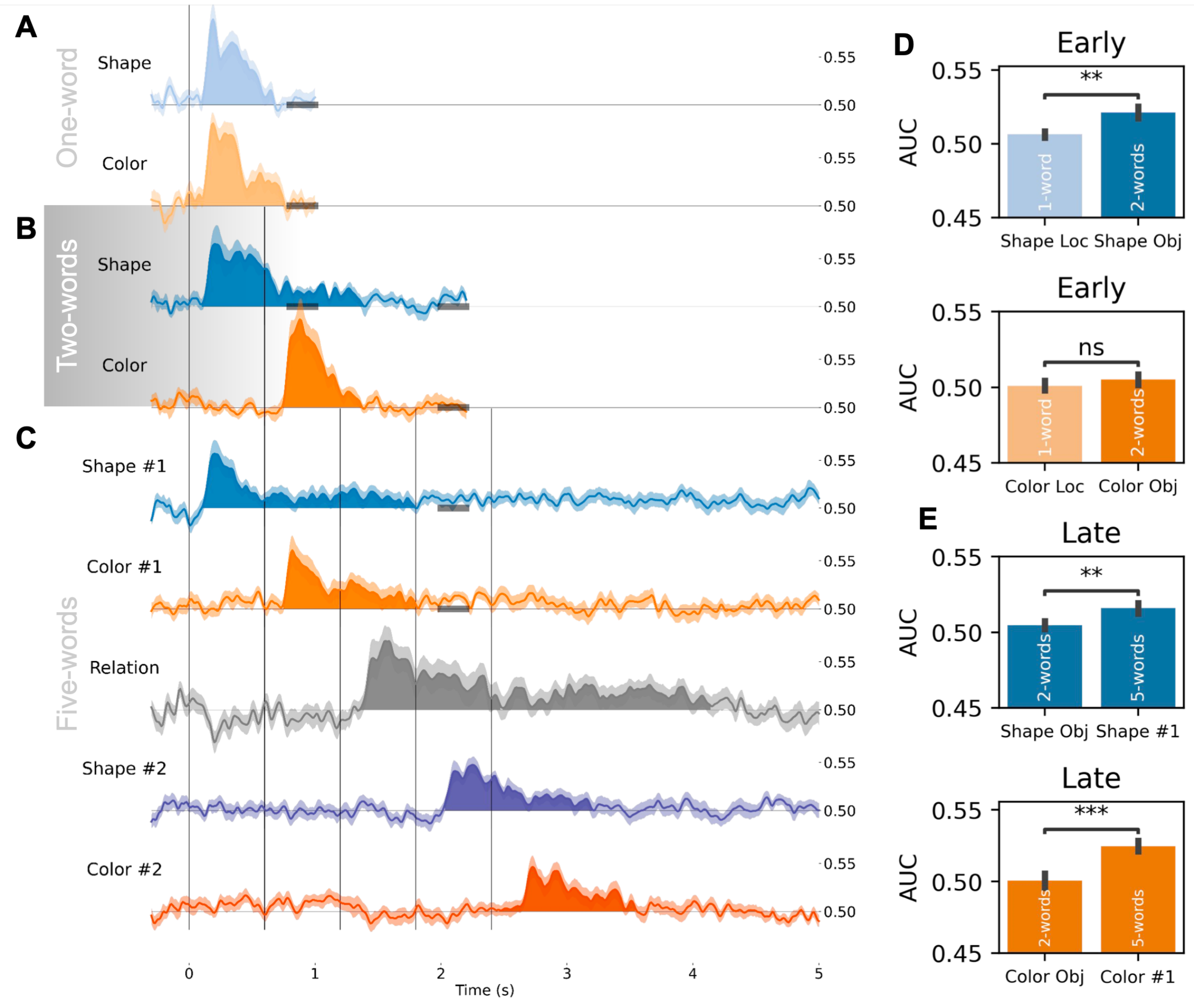


Figure 2: Single properties are actively represented until composition can occur.

A: Decoding performance over time for shape and color in one-word blocks. Shaded regions mark significant time cluster according to a permutation cluster test. The thick black lines represent the windows used for the statistical tests in D.

B: Decoding performance over time for shape and color in two-words blocks.

C: Decoding performance over time for shapes, colors, and spatial relation in five-words blocks.

D: Comparison of decoding performance between one-word blocks (left), where composition does not occur, and two-words blocks (right). All time points in the time window from 0.8 s to 1 s after word onset were fed to the classifier. The decoding performance for shape in two-words blocks is higher than the one in one-word blocks, suggesting that the representation is kept active for composition. On the other hand, using the same window for color decoding

(where in both cases no further composition should occur) does not yield any statistically significant difference. Statistics are FDR corrected.
E: Same for two-words blocks compared to five-words blocks and using a later time window (2 s to 2.2 s), where composition should be completed in two-words blocks, but not in five-words blocks. Indeed, we find that for both shape (top, blue) and color (bottom, orange), the decoding performance is higher for decoders trained on five-words blocks.

Interestingly, in two-words blocks, the shape decoding performance stayed significantly above chance (though much lower than during stimulus presentation) more-or-less as long as the color decoding performance, i.e., around 1 s after the color word onset (Figure 2B). In other words, there is an active representation of the shape while the color is being processed. Critically, decoders trained on the same words in the one-word blocks (where properties are presented individually with no composition occurring) did not exhibit this sustained decoding performance but dropped to chance-level around 700 ms after word onset (Figure 2A). This was verified by training classifiers on data from multiple time points (0.8 s to 1 s; Figure 2D); the representation of shape was still explicit for two-words blocks, but not for one-word blocks ($p<0.01$, Mann-Whitney-Wilcoxon test, FDR corrected). We replicated this analysis for color decoding, where composition should have occurred already in two-words blocks: in both one-word and two-words blocks the decoding performance stayed at chance level when trained on this time window.

Next, focus on the five-words blocks, where two shape-color pairs are presented, linked by a spatial relation. We found that the first shape and first color decoding performances stayed above chance during the whole sentence (Figure 2C). Using a later time window (from 2 s to 2.2 s), we confirmed that words were maintained later than in 2 features trials (Figure 2D, $p<0.01$ for shape, $p<0.001$ for color). Furthermore, the relation decoding performance also stayed high until around 4 s after trial onset, that is more or less 3 s after the relevant word is presented (Figure 2C, grey curve). Finally, the representation of the second shape was also maintained for some time, largely overlapping with the presentation of the second color (Figure 2C, dark blue and dark orange curves)

Later during the delay decoding performance for each property goes back to chance-level. To certify that information about the stimuli were not present in our MEG signals, we trained a strong non-linear classifier, XGBoost (Chen & Guestrin, 2016), on all time points in the late delay period (from 4 s to 5 s). Decoding performance did not exceed chance-level (e.g. for first shape decoding: mean AUC = 0.504; $p>0.05$). However, the subjects still manage to do the task with good performance (mean error rate +/- SEM: 0.042 +/- 0.006 for two-words blocks and 0.124 +/- 0.082 for five-words blocks), so this information must be stored in the brain somehow. In the next sections, we sought to identify the format of this short-term storage of language input.

During the delay period, individual features are replaced by a compressed code

If the individual properties (shape, color, and relation) are stored in a way that is not detectable in MEG signals, what parameters impact the delay activity? We hypothesized that if the sentence representation is compressed, then the amount of information (Shannon, 1948) should be reflected in the ongoing neural activity. To test this, we devised a measure of complexity (C) that quantifies the information present in our sentences as the number of non-redundant objects' properties. Put simply, if both objects share color and shape, then C = 0, if they share either color or shape then C = 1, and if they share neither shape nor color then C = 2. We thus trained a linear regression to predict the complexity of the composed representation of individual trials and evaluated the decoders with a Pearson correlation. Contrary to color and shape decoding where each class is predicted by a different classifier, this analysis learns a single mapping that predicts complexity values.

We find that the decoding performance of compositional complexity reaches significance around 2.1 s after trial onset (i.e., just before the last word's onset) and stays high until the end of the trial, while decoding performance of individual features goes back to chance-level during the delay (Figure 3A). The classifiers trained just after the presentation of the sentence generalizes poorly to later time points (Figure 3B and 3C, blue line showing the generalization of the decoder trained at 3.2 s). Then, starting around 3.8 s, decoders generalize well up to and after the image probe. For example, the purple line in figure 3C shows the generalization of the decoder trained at 4 s is above-chance when tested on time points from 3.45 s to 5.67 s, suggesting that the neural markers of complexity are stable over this duration. Finally, a peak of increased decoding performance follows the image probe, with partial generalization to earlier time points (Figure 3B and 3C, orange and yellow line), hinting that the image is translated in a format that matches the memory representation of the sentence.

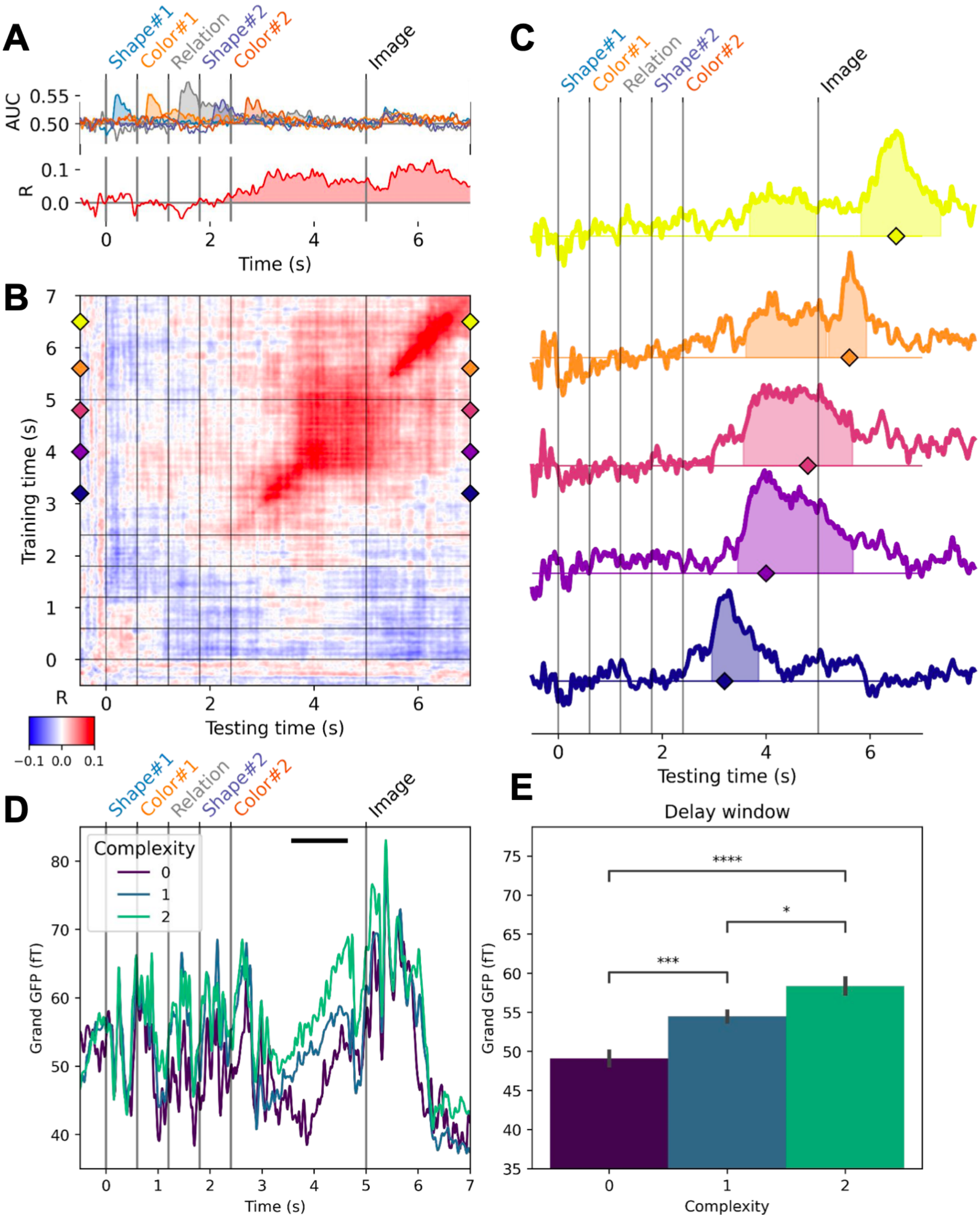


Figure 3: Neural activity in the delay period is characterized by complexity

A: Diagonal performance for decoders of individual properties (top) and complexity (bottom). The feature decoders go back to chance level during the delay, while the complexity decoding stays high.

B: Temporal generalization matrix for regression decoding of complexity

C: Horizontal slices from the temporal generalization matrix, corresponding to the generalization performance of a single classifier. Early delay period classifiers (e.g., 3 s, blue line) do not generalize well to later time points, whereas decoders trained later during the delay generalize up to and after the image probe (e.g., 4 s, purple line). Diamond markers represent the time each classifier was trained on.

D: Grand GFP on magnetometers for each level of complexity. The horizontal black bar represents the window where the statistical test is done in D.

E: Wilcoxon-Mann-Whitney test between each complexity level during the delay. The data was averaged over time points in the window and the test was performed over subjects.

These decoding results tell us that the geometry of neural signals reflect the complexity of the sentence but give no indication regarding the direction of the effect. To clarify this, we looked at the global field power (GFP) of magnetometers, averaged over subjects, for each complexity level (Figure 3D). Confirming our hypothesis, we find that during the delay the three conditions clearly diverge, with more complex trials being associated with higher GFP. We verified this using Mann-Whitney-Wilcoxon tests on GFP average over a window from 3.6 s to 4.6 s. The effect was significant for each pair of conditions (Figure 3E; complexity 0 versus 1: $p<0.001$; complexity 1 versus 2: $p<0.05$; complexity 0 versus 2: $p<0.0001$, FDR corrected)

Taken together, these results indicate that the compositional representation is compressed during the delay period. Should that be the case, how can this representation be read-out by downstream neurons to produce appropriate behavior?

Evidence for a decompression during read-out

Our results so far reveal that the memory activity is implicit (activity silent) and compressed. This final section seeks to answer the remaining question: how is information extracted back from the memory representation?

To tackle this question, we trained classifiers to differentiates trials where the image presented as a probe corresponds to the preceding sentence (match) from trials where it does not (mismatch). This gives us a window into the neural processes that foreshadow the behavioral response.

Firstly, we analyzed trials depending on their level of complexity and found that more complex trials are associated with later detection, both in neural signals (Figure 4A) and in reaction times (Figure 4B, right), but not in error rates (Figure 4C, left).

The amplitude of this shift is largest for the least complex sentences: both in neural signals and reaction times, about 200 ms separates them from the other two. On the other hand, the difference between complexity level 1 and 2 is smaller, around 50 ms, but still strongly significant ($p<0.001$ for reaction times).

Secondly, we take advantage of the hierarchical nature of our stimuli's syntactic trees, and the fact that in our experimental design, mismatch trials could be of three types. First, the *property* mismatches, where the shape or the color of one object was changed to a completely new one. Such a change in the surface properties of the syntactic tree is easily detected using a simple bag-of-words (Zhang et al., 2010). Second, the *binding* mismatches, in which either the shape or the color of the two objects are swapped. To detect this mismatch, one needs to at least parse the noun phrases of the sentences, or in other words, correctly bind the objects' shape and color (Feldman, 2013). Third, the *relation* mismatches, where individual objects are preserved, but the spatial relation that links them is reversed (e.g., "X to the left of Y" as opposed to "X to the right of Y"). Thus, detecting this error requires reaching the uppermost branch of the syntactic tree and correctly assigning the objects' location (see supplementary Figure 1).

We hypothesized that if the compositional representation is factorized then detecting each kind of mismatch should requires the same amount of time. On the other hand, if computations are needed to extract information from the memory trace, then detecting mismatches that requires a higher level of syntactic processing should take longer.

Indeed, we found that the property mismatches were detected faster than binding and relation mismatches (p<0.0001 for both; Figure 4D right) and had a lower error rate compared to relation mismatches (p<0.0001; Figure 4D left). Comparing binding and relation mismatches, we find that the reaction times do not differ significantly (p>0.05; Figure 4D, right), but that the error rate is nearly twice greater for relation mismatches (p<0.0001, Figure 4D, left). This suggests a speed-accuracy trade-off (Reed, 1973) and corroborates the subjects ’verbal reports that strongly hinted that the hardest mismatches were the relation mismatches, and the easiest the property mismatches. Interestingly, many subjects reported that for relation mismatches in particular, they answered too soon and detected very soon after that they had made a mistake.

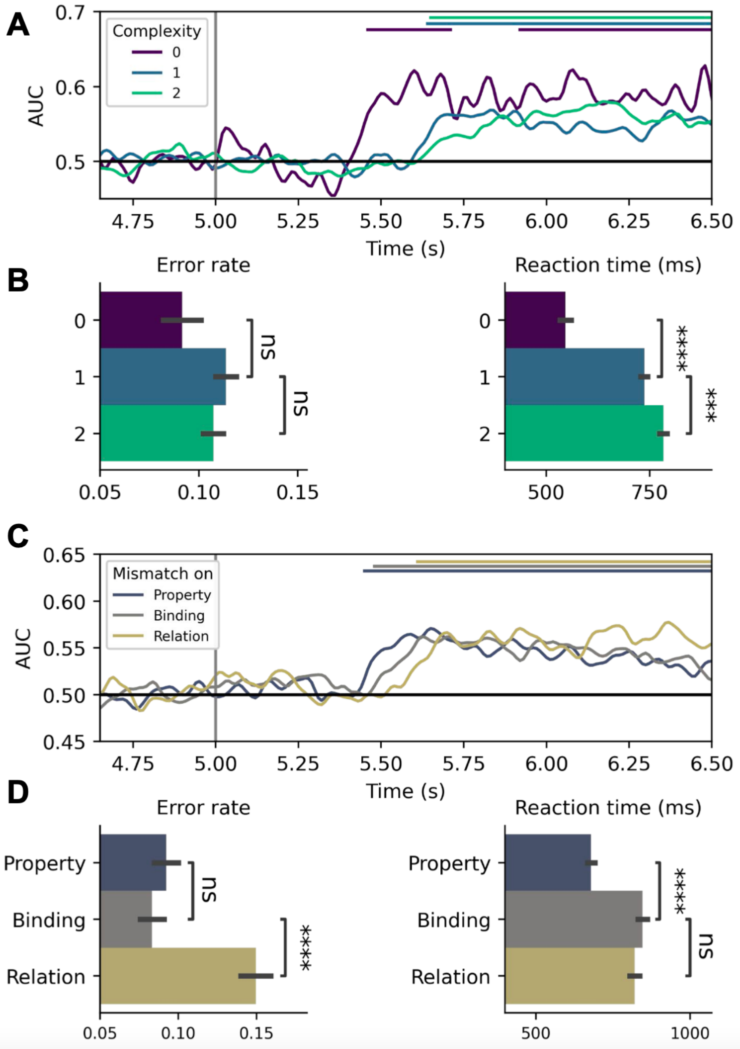


Figure 4: Read-out of composed representation is structure-dependant

A: Decoding performance of match versus mismatch trials, split for each complexity level. Lower complexity is associated with earlier segregation of match and mismatch trials in brain signals. Vertical bar marks image onset.

B: Average error rate (left) and reaction times (right) for each complexity level. The subjects' performance does not differ significantly with complexity, but their reaction time is strongly affected.
C: Decoding performance of match versus mismatch trials for each type of mismatch (Property, Binding, Relation). Mismatches touching deeper syntactic properties are associated with later segregation of match and mismatch trials in brain signals. Vertical bar marks image onset.
D: Average error rate (left) and reaction times (right) for each type of mismatch. The subjects' performance is lowest for Relation mismatches, and their reaction time is highest for Binding and Relation mismatches.

We trained three separate sets of classifiers, one for each type of mismatch, and showed their respective performance on figure 4C. This decoding analysis confirms behavioral results, with the decoder trained on match versus property mismatch starting first at 0.45 s after image onset, then the binding mismatch at 0.48 s, and the relation mismatch at 0.61 s.

Taken together, these results suggest that the maintenance of compositional representations in working memory is compressed and non-factorized, with surface properties that are apparent, but higher-level structure information needing some downstream processing to be extracted.

Discussion

We aim to track and characterize the neural representation of language composition, using the decoding of MEG activity in response to variably long sentences. Our results show that incoming words are linearly represented until they can be combined. Afterwards, these neural representations are quickly replaced by a compressed representation, whose amplitude correlates with the quantity of information in the sentence. Finally, the read-out process depends on such complexity, suggesting that a decompression operation has to take place to access properties from the stored representation. Furthermore, accessing properties that are higher in the sentence’s syntactic tree also takes longer, hinting that the properties are not stored in a factorized format, but rather need computation to be read-out.

These results complement previous studies that showed that, in English two-word phrases, the representation of the adjective is actively maintained until the processing of their associated noun is finished (Fyshe et al., 2019; Honari-Jahromi et al., 2021). We replicate this finding in French, where the word order is swapped compared to English, finding that the noun is maintained until it can be merged with its adjective. We go further, showing that the representations on the first words are kept active during most of the sentence, suggesting that this is a general property language processing in the brain. This is also compatible with the proposal that storage in working memory does not needs explicit activations, but manipulation of stored concepts does (Stokes, 2015; Trübutschek et al., 2019).

We found that colors yielded somewhat higher decoding performance, compared to shapes. This is opposite to the finding of (Honari-Jahromi et al., 2021), where noun decoding was found to be more robust than (color) adjective decoding. It may mean that, surprisingly, the second word is more easily decodable than the first.

Interestingly, the decoding performance of the relation stayed high for longer than shapes and colors (up to 4 s). This could suggest that the higher in the syntactic tree a word is, the more explicit is its neural representation after composition.

The decoding of complexity shows a rare case where neural activity during delay periods can be characterized. The first phase of increased decoding performance for complexity generalizes poorly to later timepoints and could reflect the actual compositional process. The later part of the delay is marked by a stable code, the most likely support for linguistic working memory. This finding is reminiscent of previous finding about compression in spatial working memory using primitives (Al Roumi et al., 2021; Xie et al., 2022)

These findings go against two major currents of neuroscientific theories. The first are theories of working memory, either “slot” or “resource” based (Ma et al., 2014; Bays et al., 2022) and “sustained” or “activity-silent” (Barbosa et al., 2020; Stokes et al., 2020), that do not consider manipulations to the input features before storage and at read-out. Notably, others have found that for visual working memory, working memory capacity was not impacted by complexity (Awh et al., 2016). Future work will need to untangle these contradictory results.

The second is the trend regarding factorized representations (Behrens et al., 2018; Whittington et al., 2020). Such representations have many theoretical advantages, most notably good generalization properties (Bernardi et al., 2020; Chung & Abbott, 2021). However, here we find that compositional representations are stored in an intermediate concise format and that downstream computations are needed to access the full syntactic tree, hinting that factorization happens at read-out.

The ability to compress information before storing it is necessary in a context where memory capacity is limited (Ma et al., 2014). Indeed, it has been shown that compressibility is a good predictor of working memory performance and of fluid intelligence (Chekaf et al., 2018). Here, for the first time we provide direct evidence that high-level linguistic representations are compressed in such a way, even when the load did not exceed the limits of working memory.

Note that there might be multiple alternative reasons why the properties decoding performance is at chance during the delay. First, the resolution allowed by MEG signals might not be sufficient if working memory is carried by sparse populations with sustained activity (Goldman-Rakic, 1995; Leung et al., 2002). Second, if working memory is implemented in an activity-silent manner (Stokes, 2015), there might indeed be no way to decode its content without an external “probing” signal. Third, it might be that regular “replays” of the compressed word sequence are the basis of linguistic working memory, as has been shown for other sequences (Liu et al., 2019, 2021).

Given that less complex sentences need more compression, we expected to see higher activation for less complex sentences during the early part of the delay, but this was not the case. We speculate that local signals in regions such as IFG and ATL, as would be visible with intracranial EEG, would reflect this.

Overall, we described a thorough picture of composition in our simplified setup and provide a new perspective on the nature of compositional representations in the brain.

Methods

Experimental design

In one-word blocks, the subject were asked to do a 1-back task across text and image: they were presented with a continuous stream of alternating word and image and were asked to press a button whenever the current image matched the previous word, or the current word matched the previous image (e.g., the word “circle” followed by the image of a circle, or the word “red” followed by an image of a red smudge). This setup was made to train classifiers that contain a single concept’s semantics, outside any composition operation.

In two-words and five-words blocks, the subjects were asked to read the sentence presented one word at a time and remember its content until an image appeared. Then they should press a button with their right or left hand, depending on whether the image’s content matches the sentences, or not (mismatch rate was 50%). The side of the button corresponding to “match” and “mismatch” was constant inside a block and randomized across blocks.

The experiment was split into 10 blocks, presented in a sandwich fashion: starting and ending with a one-word block, and alternating two and five-words blocks in between. The 2 one-word blocks contained 480 trials each, totaling 960 trials. This means that for each of the 6 properties (3 shapes and 3 colors), we had 960 / 6 160 trials.
The 4 two-words blocks contained 135 trials each, totaling 540 trials. This means that for each of the object’ properties (shape, color), we had 540 / 3 = 180 trials. Thus, to train a classifier to differentiate trials where, e.g., the shape was 1) “square” versus 2) “circle or triangle”, we had 180 trials in the class 1 and 380 trials in class 2.
The 4 five-words blocks contained 81 trials each, totaling 324 trials. The number of unique sentences in our design is 3 first shape * 3 first_color * 2 relation * 3 second shape * 3 second_color = 162, thus we had 2 repetitions of each unique sentence. Furthermore, for each of the marginal objects’ properties (first shape, first color, second shape, second color), we had 324 / 3 = 108 trials. Thus, to train a classifier to differentiate trials where, e.g., the first color was 1) “blue” versus 2) “red or green”, we had 108 trials in class 1 and 216 in class 2. For the spatial relationship property, we had 324 /2 = 162 trials in each category.

Regarding complexity ratings, because each feature we found with equal probability at each position, there was less trials with lower complexity (i.e., where features were identical). Specifically, we had (out of a total of 324 trials) 24 trials of complexity 0, and 150 trials for complexity 1 and 2.

In two-words blocks, there was a single king of mismatch: a new feature was selected at random to replace an existing one. E.g., “A blue square” becomes “A blue **circle***”.
Three kinds of mismatches were possible in mismatch trials (see also Supplementary Figure 1):

- In property mismatches, a new feature is selected to replace one, taken at random. This new feature could not already be present in the sentence. E.g.: “A blue circle to the left of a red square” becomes “A **green*** circle to the left of a red square”.

- In binding mismatches, two-words are swapped between the two objects. E.g.: “A blue circle to the left of a red square” becomes “A **red*** circle to the left of a **blue*** square”.
- In relation mismatches, the two objects are kept but the spatial relationship between the two is reversed. E.g.: “A blue circle to the left of a red square” becomes “A blue circle to the **right*** of a red square”.

The SOA was 600 ms for two and five-words blocks, and 900 ms for one-word blocks. The delay between last word and image onset was 1 s for two-words blocks and 2 s for five-words blocks. The image was kept on screen for 600 ms, then a response screen reminded the participant which button corresponded to “match” and “mismatch”. Because this mapping was constant inside a block, subjects were asked to answer as fast as possible, not necessarily waiting for the response screen,

Multivariate decoding

For each object’s properties (shape and color), we have 3 classes (“red”, “green”, “blue” for colors and “circle”, “square”, “triangle” for shapes). At each time point in MEG single-trial data, we trained a logistic regression to separate each of these properties in a One-Versus-Rest fashion, meaning that each class was tested against the two other classes. e.g., “red” was tested against “green and blue”. The decoding performance reported is the averaged over the 3 classifiers for each property. Decoding the spatial relationship is a simple binary classification problem.

Such a decoding analysis informs us about whether and when our experimental conditions are differently represented in neural signals: if at time t the classifier reaches above-chance performance, it means that the brain signals contain information about the shape or color at this time. If the decoding performance is at chance, it means that the signals do not contain any such information, either because it is not present in the brain (e.g., before the trial starts), or because it is not represented in a way that can be detected with MEG recordings (e.g., during the delay). These classifiers were then tested at each other time point according to the temporal generalization method (King & Dehaene, 2014). This extension of the traditional within-time decoding analysis allows to test for the consistency of neural patterns over time: if a classifier trained at time t generalizes to time T, it means that the neural patterns is somewhat similar between time t and T. On the other hand, within-time decoding could be high at both t and T, but with no generalization between t and T. This would mean that the brain segregates stimuli at both time points, but with a different pattern of activations. In other words, the within-time decoding performance (trained and tested at the same time, i.e., the diagonal of the temporal generalization matrix) inform us about the content of brain signals, while the across-time decoding performance (trained and test at different times, i.e., the off-diagonal elements) tells us about the stability of these representations.

For decoding in one-word blocks, only trials where a word (not an image) was presented were used to train the classifier. This was done to be fully comparable to two and five-words blocks. Moreover, at test time, only trials that were not followed by a matching image were used, because a matching image would have confounded that memory trace with the incoming stimulus.

For the regression decoding of complexity, the score was computed using a Pearson correlation between the (cross-validated) predicted and actual complexity. With this setup, to reach good decoding performance the three complexity levels need not only to be linearly separable, but also to respect the ordering we specified (0 < 1 < 2).

Before training each classifier, the data was subtracted from its median and scaled using the interquartile range, i.e. the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile). We used a stratified 10-fold cross validation procedure. We average the classifiers' performances across all folds and report the average performance across subjects. All neural data analyses were performed using MNE-Python (Gramfort et al., 2013) and scikit-learn (Pedregosa et al., 2011).

Global Field Power (GFP)

GFP is an aggregate measure commonly used in electrophysiological data (Michel et al., 1993). It combines information from all channels by computing their standard deviation. Here, we compute the GFP for each subject and report the average.

Bibliography

Al Roumi, F., Marti, S., Wang, L., Amalric, M., & Dehaene, S. (2021). Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. *Neuron*, *109*(16), 2627-2639.e4. https://doi.org/10.1016/j.neuron.2021.06.009

Awh, E., Barton, B., & Vogel, E. K. (2016). Visual Working Memory Represents a Fixed Number of Items Regardless of Complexity. *Psychological Science*. https://journals.sagepub.com/doi/10.1111/j.1467-9280.2007.01949.x

Barbosa, J., Stein, H., Martinez, R. L., Galan-Gadea, A., Li, S., Dalmau, J., Adam, K. C. S., Valls-Solé, J., Constantinidis, C., & Compte, A. (2020). Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nature Neuroscience*, *23*(8), Article 8. https://doi.org/10.1038/s41593-020-0644-4

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190307. https://doi.org/10.1098/rstb.2019.0307

Bays, P., Schneegans, S., Ma, W. J., & Brady, T. (2022). *Representation and computation in working memory*.

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

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

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, S0092867420312289. https://doi.org/10.1016/j.cell.2020.09.031

Blouw, P., Solodkin, E., Thagard, P., & Eliasmith, C. (2016). Concepts as Semantic Pointers: A Framework and Computational Model. *Cognitive Science*, *40*(5), 1128–1162. https://doi.org/10.1111/cogs.12265

Brennan, J., & Pylkkänen, L. (2012). The time-course and spatial distribution of brain activity associated with sentence processing. *NeuroImage*, *60*(2), 1139–1148. https://doi.org/10.1016/j.neuroimage.2012.01.030

Brennan, J. R., & Pylkkänen, L. (2017). MEG Evidence for Incremental Sentence Composition in the Anterior Temporal Lobe. *Cognitive Science*, *41*(S6), 1515–1531. https://doi.org/10.1111/cogs.12445

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv:2005.14165 [Cs]*. http://arxiv.org/abs/2005.14165

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and Generalization In Emergent Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4427–4442.

Chekaf, M., Gauvrit, N., Guida, A., & Mathy, F. (2018). Compression in Working Memory and Its Relationship With Fluid Intelligence. *Cognitive Science*, *42*(S3), 904–922. https://doi.org/10.1111/cogs.12601

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62. https://doi.org/10.1017/S0140525X1500031X

Chung, S., & Abbott, L. F. (2021). Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, *70*, 137–144. https://doi.org/10.1016/j.conb.2021.10.010

Conwell, C., & Ullman, T. (2022). *Testing Relational Understanding in Text-Guided Image Generation* (arXiv:2208.00005). arXiv. https://doi.org/10.48550/arXiv.2208.00005

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A Large-Scale Model of the Functioning Brain. *Science*, *338*(6111), 1202–1205. https://doi.org/10.1126/science.1225266

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, *113*(41), E6256–E6262. https://doi.org/10.1073/pnas.1612132113

Feldman, J. (2013). The neural binding problem(s). *Cognitive Neurodynamics*, *7*(1), 1–11. https://doi.org/10.1007/s11571-012-9219-8

Frankland, S. M., & Greene, J. D. (2020). Concepts and Compositionality: In Search of the Brain’s Language of Thought. *Annual Review of Psychology*, *71*(1), 273–303. https://doi.org/10.1146/annurev-psych-122216-011829

Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A., & Bolhuis, J. J. (2017). Language, mind and brain. *Nature Human Behaviour*, *1*(10), Article 10. https://doi.org/10.1038/s41562-017-0184-4

Friederici, A. D., Kotz, S. A., Scott, S. K., & Obleser, J. (2010). Disentangling syntax and intelligibility in auditory language comprehension. *Human Brain Mapping*, *31*(3), 448–457.

Fyshe, A., Sudre, G., Wehbe, L., Rafidi, N., & Mitchell, T. M. (2019). The lexical semantics of adjective–noun phrases in the human brain. *Human Brain Mapping*, *40*(15), 4457–4469.

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, *9*(9), 416–423. https://doi.org/10.1016/j.tics.2005.07.004

Honari-Jahromi, M., Chouinard, B., Blanco-Elorrieta, E., Pylkkänen, L., & Fyshe, A. (2021). Neural representation of words within phrases: Temporal evolution of coloradjectives and object-nouns during simple composition. *PLOS ONE*, *16*(3), e0242754. https://doi.org/10.1371/journal.pone.0242754

Humphries, C., Love, T., Swinney, D., & Hickok, G. (2005). Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Human Brain Mapping*, *26*(2), 128–138.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198270126.001.0001

Kajić, I., Schröder, T., Stewart, T. C., & Thagard, P. (2019). The semantic pointer theory of emotion: Integrating physiology, appraisal, and construction. *Cognitive Systems Research*, *58*, 35–53. https://doi.org/10.1016/j.cogsys.2019.04.007

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

Leung, H.-C., Gore, J. C., & Goldman-Rakic, P. S. (2002). Sustained Mnemonic Response in the Human Middle Frontal Gyrus during On-Line Storage of Spatial Memoranda. *Journal of Cognitive Neuroscience*, *14*(4), 659–671. https://doi.org/10.1162/08989290260045882

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human Replay Spontaneously Reorganizes Experience. *Cell*, *178*(3), 640-652.e14. https://doi.org/10.1016/j.cell.2019.06.012

Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D., & Dolan, R. J. (2021). Experience replay is associated with efficient nonlocal learning. *Science*, *372*(6544). https://doi.org/10.1126/science.abf1357

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), Article 3. https://doi.org/10.1038/nn.3655

Marcus, G., Davis, E., & Aaronson, S. (2022). *A very preliminary analysis of DALL-E 2* (arXiv:2204.13807). arXiv. http://arxiv.org/abs/2204.13807

Martin, A. E., & Doumas, L. A. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLOS Biology*, *15*(3), e2000663. https://doi.org/10.1371/journal.pbio.2000663

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L., & Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, *5*(4), 467–479. https://doi.org/10.1162/jocn.1993.5.4.467

McCoy, R. T., Linzen, T., Dunbar, E., & Smolensky, P. (2019). *RNNs Implicitly Implement Tensor Product Representations* (arXiv:1812.08718). arXiv. https://doi.org/10.48550/arXiv.1812.08718

Michel, C. M., Brandeis, D., Skrandies, W., Pascual, R., Strik, W. K., Dierks, T., Hamburger, H. L., & Karniski, W. (1993). Global field power: A ‘time-honoured’ index for EEG/EP map analysis. *International Journal of Psychophysiology*, *15*(1), 1–2. https://doi.org/10.1016/0167-8760(93)90088-7

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science*, *319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769

Nelson, M. J., Karoui, I. E., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., & Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National*

Academy of Sciences, *114*(18), E3669–E3678. https://doi.org/10.1073/pnas.1701590114

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527. https://doi.org/10.1073/pnas.1018711108

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Planton, S., van Kerkoerle, T., Abbih, L., Maheu, M., Meyniel, F., Sigman, M., Wang, L., Figueira, S., Romano, S., & Dehaene, S. (2021). A theory of memory for binary sequences: Evidence for a mental compression algorithm in humans. *PLoS Computational Biology*, *17*(1), e1008598. https://doi.org/10.1371/journal.pcbi.1008598

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, *366*(6461), 62–66. https://doi.org/10.1126/science.aax0050

Pylkkänen, L. (2020). Neural basis of basic composition: What we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190299. https://doi.org/10.1098/rstb.2019.0299

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International*

Conference on Machine Learning, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *ArXiv Preprint ArXiv:2204.06125*.

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science*, *181*(4099), 574–576.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review*, *77*, 481–495. https://doi.org/10.1037/h0029964

Restle, F., & Brown, E. R. (1970). Serial pattern learning. *Journal of Experimental Psychology*, *83*, 120–125. https://doi.org/10.1037/h0028530

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., & others. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *ArXiv Preprint ArXiv:2205.11487*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216. https://doi.org/10.1016/0004-3702(90)90007-M

Soulos, P., McCoy, T., Linzen, T., & Smolensky, P. (2020). *Discovering the Compositional Structure of Vector Representations with Role Learning Networks* (arXiv:1910.09113). arXiv. http://arxiv.org/abs/1910.09113

Spaak, E., Watanabe, K., Funahashi, S., & Stokes, M. G. (2017). Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. *The Journal of Neuroscience*, *37*(27), 6503–6516. https://doi.org/10.1523/JNEUROSCI.3364-16.2017

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Stokes, M. G., Muhle-Karbe, P. S., & Myers, N. E. (2020). Theoretical distinction between functional states in working memory and their corresponding neural states. *Visual Cognition*, *28*(5–8), 420–432. https://doi.org/10.1080/13506285.2020.1825141

Thagard, P., & Stewart, T. C. (2014). Two theories of consciousness: Semantic pointer competition vs. information integration. *Consciousness and Cognition*, *30*, 73–90. https://doi.org/10.1016/j.concog.2014.07.001

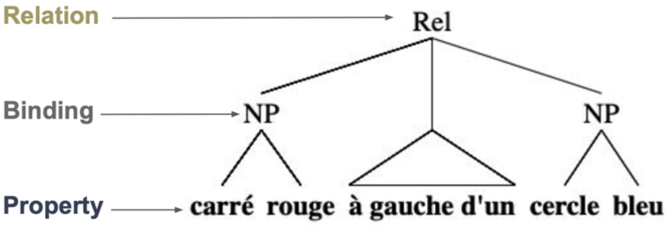
Trübutschek, D., Marti, S., Ueberschär, H., & Dehaene, S. (2019). Probing the limits of activity-silent non-conscious working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(28), 14358–14367. https://doi.org/10.1073/pnas.1820730116

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, *183*(5), 1249-1263.e23. https://doi.org/10.1016/j.cell.2020.10.024

Xie, Y., Hu, P., Li, J., Chen, J., Song, W., Wang, X.-J., Yang, T., Dehaene, S., Tang, S., Min, B., & Wang, L. (2022). Geometry of sequence working memory in macaque prefrontal cortex. *Science*, *375*(6581), 632–639. https://doi.org/10.1126/science.abm0204

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1), 43–52.

Supplements



Supplementary Figure 1: Syntactic depth of each kind of mismatch

The three types of mismatches affect properties of the syntactic tree at various depths. The property mismatches impact only surface properties. Binding mismatches affects the formations of noun phrases (NP), while relation mismatches reach the highest level.