# Disentangling Hierarchical and Sequential Computations during Sentence Processing

**Christos-Nikolaos Zacharopoulos**[1*], **Stanislas Dehaene**[1,2], **and Yair Lakretz**[1*]

[1]Cognitive Neuroimaging Unit, NeuroSpin center, 91191, Gif-sur-Yvette, France
[2]College de France, 11 Place Marcelin Berthelot, 75005 Paris, France

**Sequence processing in humans is thought to rely on two distinct mechanisms: the computation of transition probabilities between adjacent elements and the extraction of larger hierarchical structures. Previous studies indicate that both mechanisms contribute to auditory sequence processing, but whether language processing involves one or the other remains debated. To address this issue, we designed a linguistic version of the local-global auditory test, which contrasts sequential versus hierarchical violations of expectations in sentences, and we searched for violation responses in both human magnetoencephalography and computational models. We found that in models, both mechanisms coexist, whereas humans only show hierarchical structure effects. Our results suggest that human sentence processing is dominated by structure-based computations and robust to sequential effects. They point to major differences between language processing in humans versus neural models and, within humans, between the processing of linguistic sentences versus non-linguistic sequences.**

**Correspondence:** ***christonik@gmail.com (CNZ), yair.lakretz@gmail.com (YL)***

## Introduction

A central question in linguistics pertains to the nature of linguistic structures and their processing. According to formal linguistics, language processing is enabled via a largely innate and human-specific ability to mentally transform and represent sequentially incoming words into a hierarchical tree structure (1–3). Others, however, claim that sequential-sentence structure has considerable explanatory power and that hierarchical processing is often not involved in language use (4). According to this view, language ability does not rely on an innate ability, and it is rather acquired through a process in which children gradually build up a repertoire of increasingly abstract constructions (5, 6).

Fundamental to the hierarchical view of sentence structures is the idea that structures of sentences are made up of *constituents* – groups of words that belong together and are nested within each other. For example, the constituent structure of the sentence 'The boy near the girls likes climbing' can be described by a tree-structure representation (Figure 1A), or by the equivalent bracket notation [[[The boy] [near [the girls]]] [likes climbing]]]. The hypothesis of a tree-based constituent structure is supported by considerable evidence, including substitutions: some groups of words can be removed or replaced by a shorter expression, but only if they form a valid constituent. For example, the nested constituent [the girls] can be replaced by 'them' - [[[The boy] [near them]] [likes climbing]], or it can be removed from the sentence altogether. Similarly, the larger constituent [The boy near the girls] can be replaced by 'He' - [[He] [Likes climbing]]. Importantly, this is not the case for any block of words in the sentence. For example, 'girls likes' cannot be removed or replaced by another word. Additional support for hierarchical structures comes from long-range agreements in language. For example, in the above sentence, the main subject 'boy' and the verb 'likes' agree on their grammatical number (singular) despite their separation and the presence of an intervening noun 'girls' that carries the opposite grammatical number (plural). To explain such long-range agreements, the underlying latent structure of the sentence needs to be hierarchical, linking 'boy' and 'likes' despite their temporal separation (note that in the tree-structure representation in Figure 1A, the distance between the nodes for 'boy' and 'likes' is shorter than that between 'girls' and 'likes'). The hierarchical view is further supported by a large number of linguistic analyses, based on, e.g., binding (7), question construction (8), prosodic prominence in morphology (9), phonology (10, 11), language acquisition (12, 13), as well as from behavioral (14) and brain-imaging experiments (15–22)

In contrast to the hierarchical view, it was argued that the sequential structure of sentences is more fundamental to language use than hierarchical structures (4). According to this view, sentence processing is dominated by predictions that are learned from statistical regularities among sequences of linguistic items, rather than from hierarchical relations among them. Hierarchical structures, it is agreed, may play a role in descriptions of language, but for language use, sequential structures are thought to have primacy. As an example of a sequential structure, the human cognitive system might process the above sentence as: [The boy] [near the girls] [likes climbing]. In this case, words are combined into chunks with sequential order but no further nested structure. Such a sequential analysis of sentence structures is therefore more parsimonious than a hierarchical one – sentences trivially have sequential structure, whereas hierarchical structure is only revealed through linguistic analyses. Since the sequential view pertains to language use only, support for this view has mainly relied on empirical data from language acquisition (23), behavioral (24–27) and brain-imaging experiments (28–30).

The hierarchical and sequential views of sentence structures are often seen as mutually exclusive. Here, however, we ask whether both sequential and hierarchical processing co-exist

in language use. For sequences of non-linguistic auditory items, it was indeed shown that multiple types of processes can co-exist, each operating on a different level of encoding of incoming sequences. Specifically, using the Local-Global Paradigm, a variant of the oddball auditory paradigm, Bekinschtein et al. (31) showed that transition-based (local) processing can be distinguished from that of chunking (global) and from possibly higher levels: in a variety of brain-imaging experiments, using EEG, MEG, intracranial recording and functional MRI, those two processes elicit distinct violation signals which differ in latency, brain areas, time-frequency signatures, and resistance to distraction or loss of consciousness (32–35). Extrapolating these findings to language, distinct neural mechanisms might also underlie sequential and hierarchical processing of sentences. One mechanism would be sensitive to local statistical regularities among word sequences, and the other to latent hierarchical relationships.

Moreover, for language, recent studies on artificial neural language models have shown that such mechanisms do coexist and that they emerge during training without explicit supervision (36, 37). Although language models are typically trained to predict the next word (a local transition problem), they can develop abstract, complex multi-word representations to do so. One type of mechanism was shown to be sensitive to sentence structure, and to be carried by a few dedicated units in the network termed 'Syntax units' and 'Long-range number units'. The second type of mechanism was shown to be sensitive to local word transitions in language and to be carried by a larger set of units, termed 'Short-range' units. The short-range units carry predictions about upcoming words based on surface statistical regularities among word sequences, whereas the long-range units generate predictions based on the hierarchical structure of the sentence. Predictions in the models about upcoming words are therefore composed of processes occurring at distinct levels of encoding of incoming sequences.

Suppose now that the human brain also contains such two distinct types of neurons, in low- and high-level regions. During sentence processing, we might then observe neural activity arising from these distinct neural processes, similarly to the findings in models and during non-linguistic sequence processing in humans. To test this hypothesis, we created a new design, akin to the local-global paradigm, which directly contrasts sequential and hierarchical computations during sentence processing. Like the local-global paradigm, it uses a 2x2 design in which both local (based on transition probabilities) and long-distance relations of agreement are manipulated orthogonally and can be independently respected or violated. We recorded neural activity from both human participants ($n = 22$, magnetoencephalography; MEG) and a neural language model and studied whether the neural signatures of the two levels can be disentangled. We used temporally resolved decoding techniques (34), which we applied to neural data from both humans and a neural language model. To anticipate on the results, in the model, we found evidence for two distinct neural effects, which correspond to the two types of processing, corroborating previous results in NLMs. In human brain signals, however, only main effects related to the hierarchical structure of sentences were found, whereas decoding of sequential effects remained at chance level. This suggests that unlike low-level processing of auditory stimuli, in sentence processing, once words enter the language system, computations are dominated by structure-based processing and are largely robust to sequential effects.

**A. The Local-Global Paradigm for Sentence Processing.** We start with a short description of the classic "local-global paradigm", whose design is parallel to that suggested here for sentence processing. The local-global paradigm is a variant of the auditory oddball paradigm (31), in which participants are presented with sounds in short sequences, instead of in a continuous manner as in the classic oddball paradigm, which allows them to perform chunking. The local-global paradigm shows that when participants are presented with $aaaab$ sequence patterns, in which the first four tones are identical and the fifth differs, the deviancy of the last tone generates a mismatch response (MMR) followed by a late surprise-elicited $P3b$ wave (31). A repetition of the same $aaaab$ sequence reduces the $P3b$ component, however, the "local" effect of MMR remains, suggesting that the MMR is an automatic response to local transition probabilities. The disappearance of the $P3b$ component suggests that a "global" expectation for a deviant fifth tone was generated. Indeed, when a $aaaaa$ pattern is subsequently presented, the $P3b$ wave reappears, showing that a monotonic sequence can be surprising if it violates prior expectations (38). The neural signatures of the local and global effects further differ in several ways. First, the local effect is early (100–200ms) and transient, whereas the global effect requires additional 100-200 ms to rise, and it remains stable (34). Second, the local effect is automatic (does not require attention) and unconscious, whereas the global one disappears when participants are not attending or unconscious (31, 35). Third, while the local effect was traced back to auditory cortices (32, 39), the global one is distributed across the superior temporal sulcus, inferior frontal gyrus, dorsolateral prefrontal, intraparietal, anterior, and posterior cingulate cortices (40). Taken together, this shows that there exist two distinct neural mechanisms involved in the processing of sequence patterns, with very different properties and which are sensitive to regularities at different scales.

Extrapolating from sequences of simple tones to words, we hypothesize that distinct neural mechanisms might also underlie sentence processing. For sentence processing, the local effect would correspond to local transition between adjacent words in a sentence, whereas the global effect would correspond to structural relations among distant words, where one word is syntactically dependent on another one. To contrast syntactic and local expectation, we make use of grammatical agreement, which is considered one of the best proxies to syntactic computations during sentence processing (41). This is since grammatical agreement is ruled by the latent structure of the sentence rather than by its linear (sequential) order of words. Consider for example the following sentence, which contains a prepositional phrase (PP; Figure 1A):

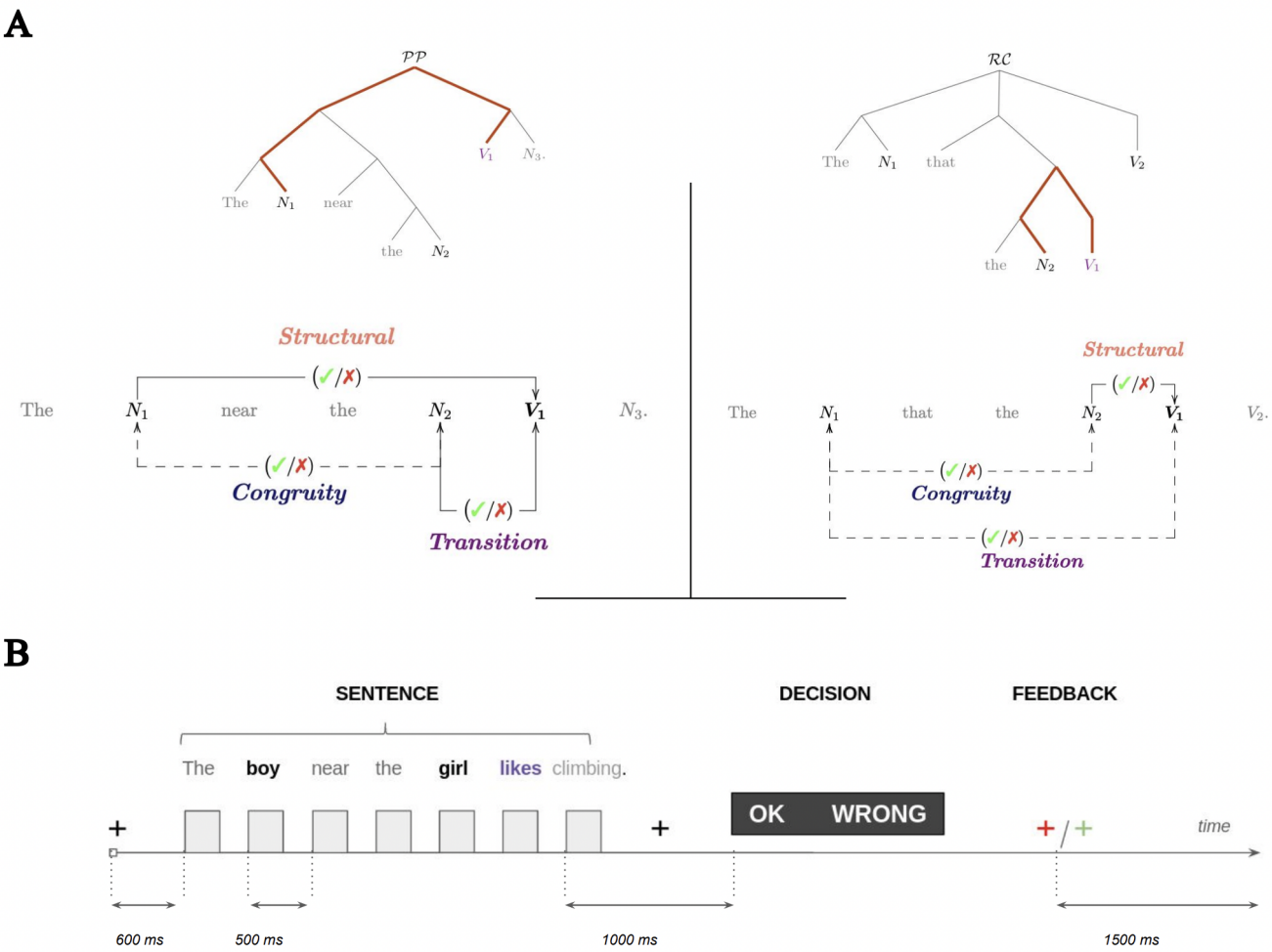


**Fig. 1. Structural vs. linear intervention in sentence processing—experimental design and paradigm.** To disentangle two possible types of processing during sentence comprehension, the experimental design contrast: (i) a structural dependency between a target verb and a noun, which either holds, or creates a violation at the verb, and (ii) a linear (sequential) interaction between the target verb and another noun, which either facilitates or interferes with verb processing. (A) Tree representations of the two sentence constructions explored in the experiments. Below, an illustration of the main effects of the design: Structural effect (orange), which depends on the syntactic relation between the main subject and target verb (colored path in the tree representation). Transition effect (magenta), which refers to the (mis)match between the target verb and a linearly intervening noun (attractor), with respect to either grammatical number or animacy. congruity effect, which refers to the (mis)match between the two nouns; In the left construction, the structural effect is long-range and the transition (linear) one is short-range. On the right construction, it is the opposite. (B) Experimental Paradigm: subjects were presented with sentences in a rapid serial visual presentation (RSVP), and their task was to report whether the sentences are grammatically correct. At the end of each trial, a visual feedback on their performance was given.

(1) **PP**: "The **boy** near the girls **likes** climbing" ("$Det\ N_1\ P\ det\ N_2\ V\ N_3$").

In this sentence, the structural dependency between the main subject 'boy' $(N_1)$ and verb 'likes' $(V)$ requires that they agree on the grammatical number (singular), despite their linear separation. In contrast, the second noun, 'girls' $(N_2)$, is adjacent to the verb but does not stand in structural relations with it. It rather linearly intervenes between the main subject and verb, possibly generating (erroneous) local expectations for a plural form of the upcoming word. Transition- and structure-based processing can be therefore separately manipulated by:

1. The structural relation between $N_1$ and $V$ (Global manipulation).
2. The linear intervention of $N_2$ on the processing of $V$ (Local manipulation).

These two dimensions span the two-by-two design of the paradigm, and its corresponding two main effects are defined as follows (Figure 1): A structural effect, which contrasts conditions in which $N_1$ and $V$ agree and disagree on grammatical number. In the case of disagreement, a syntactic violation occurs, and a neural response to this violation is expected, as was extensively studied in past studies (42, 43). The second one is a transition effect, which contrasts conditions in which $N_2$ and $V$ match and mismatch with respect to grammatical number. The transition effect corresponds to local word transitions and was not identified in neural recordings thus far. Following results from the classic local-global paradigm and from simulations in neural language models, we hypothesized that a local number mismatch would violate local word-transition expectations, which, in turn, would generate an identifiable neural response, independently of whether a syntactic violation simultaneously occurs. For ex-

| Construction | Template | Examples | *Violation* | *Congruity* |
|---|---|---|---|---|
| $PP - Number$ | Det $N_1$ P det $N_2$ $V_1$ $N_3$ | The boy near the girl likes climbing. | No | Yes |
| | | The boy near the girls likes climbing. | No | No |
| | | The boy near the girls like climbing. | Yes | No |
| | | The boy near the girl like climbing. | Yes | Yes |
| $ObjRC - Number$ | Det $N_1$ Adv det $N_2$ $V_1$ $V_2$ | The boy that the girl likes leaves. | No | Yes |
| | | The boy that the girls like leaves. | No | No |
| | | The boy that the girls likes leaves. | Yes | No |
| | | The boy that the girl like leaves. | Yes | Yes |
| $PP - Animacy$ | Det $N_1$ P det $N_2$ $V_1$ $N_3$ | The boy near the girl likes climbing. | No | Yes |
| | | The boy near the car likes climbing. | No | No |
| | | The boy near the car rusts badly. | Yes | No |
| | | The boy near the girl rusts badly. | Yes | Yes |

**Table 1. Design and prototypical examples.** The experiment utilizes two linguistic constructions, a Prepositional Phrase (PP) and an Object Relative Clause (ObjRC) as well as two features of interest (Number & Animacy). The manipulation of the violation factor corresponds to the structural effect.

ample, in sentence (1), the frequency of 'girls likes' is two orders of magnitude smaller than that of 'girl likes' (log-frequency = -8 and -6, respectively; Google's ngram). Low-level brain regions of the language network might be sensitive to such transition probabilities, in which case, a greater neural response is predicted for the low- compared to the high-frequency word pair. This is based on a predictive-coding framework (44–46), which suggests that cortical circuits form an internal model of input sequences, and that this model continuously generates predictions about upcoming items, confronting them with incoming stimuli. The local effect would thus reflect a prediction error that results from an internal model based on transition probabilities, in accordance with the 'sequential' view described above (25).

**B. Decoupling syntactic dependency and linear proximity.** Notice that in the construction with a PP, the global (structural) and local (transition-based) effects are correlated with linear proximity. That is, the global effect between the subject $N_1$ and verb $V$ is long-range, whereas the local effect between the intervening noun $N_2$ and verb is short-range. To decouple syntactic dependency and linear proximity, we included a control construction in the design, in which syntactic dependency and linear proximity are reversed by the replacement of only a single word. Specifically, we replaced the PP by an object-relative clause (ObjRC; Figure 1, Table 1):

(2) **ObjRC**: "The boy that the **girls like**…" ("$Det\ N_1\ that\ det\ N_2\ V$"),

In this case, the structural dependency (in bold) is now between $N_2$ and $V$, whereas the intervening noun is the distant $N_1$. The difference between the two sentences (1) and (2) is minimal – only at the third word ('near'/'that'), and the number of words that precede the target verb $V$ is the same. This allows us to test the impact of the length of the subject-verb dependency on the global effect, and the impact of the proximity between $N_2$ and $V$ on the local effect. In particular, to test the prediction that a neural response to local-transition violations would occur only in the case of a PP but not in the ObjRC case.

**C. Local and global effects for animacy violations.** So far, the examples shown contained variations of a sentence with respect to the feature of grammatical number. The number feature and the corresponding agreement phenomena are generally perceived as a proxy into syntactic processing. However, violation responses are known to vary depending on whether the violation is semantic or syntactic. While syntactic violations typically generate a late positive neural response P600 (47), semantic violations were found to elicit an earlier negative response N400 (48). The local-global design therefore further manipulates the type of feature in the agreements, and it includes sentences of the form (1), in which violations are with respect to animacy, for example:

(3) **PP**: "The **boy** near the car **likes** climbing" ("$Det\ N_1\ near\ det\ N_2\ V$"),

In these sentences, the subject $N_1$ and verb $V$ always agree on number, however, the sentence contain a semantic violation, which is either local ('car likes'), or global (as in, "The boy near the girl rusts badly"). Table 1 summarizes the three constructions of the design along with example sentences. Last, note that due to a time constraint on the entire experimental duration, we did not include a fourth case in the design, which includes sentences with an object-relative clause and semantic violations.

## Results

We first present the behavioral results from the experiment, based on the performance of the subjects in a forced-choice, violation-detection task (Figure 1C). We then present classification results in a time-resolved manner on the main effects of the design, for both humans (MEG & EEG) and artificial language models.

**D. Behavioral Results.** Figure 2 shows the mean error rates across all participants for the three main constructions. We present the error-rates with respect to the structural and congruity effects, where by congruity effect we refer to whether the two nouns $N_1$ and $N_2$ agree on number or not.

We first tested whether the *structural effect* (Figure 1C) is present at the behavioral level. Indeed, this effect was significant across all constructions. This indicates that participants made more errors in detecting a violation, compared to an affirming that a sentence was grammatical, which is akin to 'grammatical illusion' (49). The effect was stronger for PP-Number (Violation: $22.1 \pm 1.78\%$; Grammatical:$7.53 \pm 0.92\%$; $F(21) = 56.94$, $p < 1e-10$), followed by PP-Animacy (Violation: $12.63 \pm 1.2\%$; Grammatical: $6.66 \pm 0.81\%$; $F(21) = 16.85$, $p < 1e-04$) and finally ObjRC-Number (Violation: $22.86 \pm 2.01\%$; Grammatical:$14.52 \pm 1.32\%$; $F(21) = 13.25$, $p < 1e-03$).

We then examined the effect of *congruity* on error rate. Notably, this effect was significant only for the number feature, that is, we did not observe any influence of the embedded noun in the processing of the subject-verb agreement when the intervening noun mismatched the main noun with respect to animacy (PP-Number: $F(21) = 11.77$, $p < 1e-03$, ObjRC $F(21) = 15.67$, $p < 1e-03$).

Finally, we investigated the *transition effect*, which corresponds to a mismatch between the intervening noun and the target verb, e.g., $N_1$ and $V$ in (2). Notably, as with the congruity effect, this effect was significant only in the case of the number feature (PP-Number: $F(21) = 3.99, p = 0.04$, ObjRC-Number: $F(21) = 6.2, p = 0.01$). The transition effect for number is consistent with 'grammatical asymmetry' (50), where participants make more agreement errors on ungrammatical compared to grammatical sentences.

Finally, we detected a significant construction effect. Overall, participants made more errors on sentences with an object relative clause compared to a prepositional phrase ($PP$ : $3.05 \pm 2.9\%$; $ObjRC = 4.64 \pm 3.31\%$, $F(21) = 15.82$, $p < 1e-03$).

In summary, at the behavioral level we observed a significant structural effect, for both constructions (PP & ObjRC) and both features (grammatical number and animacy). Transition and congruity effects were significant for both constructions but for grammatical number only. And, finally, participants made more errors on agreement in an embedded clause compared to across a prepositional phrase.

**E. Structural but not Transition Effects are Decodable in Human Data..** The transition and congruity effects for grammatical number in the behavioral data suggest that the representation of the two nouns, $N_1$ and $N_2$, interact during sentence processing. There could be two types of interactions: (1) Interference in memory. An incongruent number of $N_2$ interferes with that of $N_1$ in memory; (2) Conflicting prediction signals. A (global) expectation for the corresponding grammatical number of $V$, based on $N_1's$ number, is affected by a (local) expectation generated by $N_2$. However, the behavioral results cannot tell apart the two, since both alternatives predict the same behavioral outcome. Therefore, we next tested whether these effects are traceable in the neural data.

We sought for the effects in both the model and in humans. We presented the same stimuli to human participants and to the model, and recorded network activity after the presentation of each word. For humans, neural activity was recorded with a magnetoencephalography (MEG) machine. For the model, we extracted hidden activity of all recurrent units of the network (Methods).

To identify the main effects in the data, we used standard decoding techniques: for each effect, at each time point, a linear binary classifier was trained to separate trials from the two conditions, and then tested on unseen data in a cross-validation manner. Figure 3 shows the decodability of the main effects for both the artificial (panel A) and human data (panel B).

For the model (Figure 3A), for all three constructions, all effects were decodable with high performance, measured in terms of the Area Under of Curve (AUC). The structural effect reached full decodability after the onset of the target word. Indeed, prior to the onset of the target verb, the model cannot predict the grammaticality of the sentence. The transition effect reached full decodability also after the onset of the target word. Here too, prior to verb onset, a mismatch between the verb and the non-head noun cannot be predicted. Finally, the congruity effect was decodable already after the onset of the second noun. Indeed, information about feature mismatch between the two nouns is already available at this time point.

For the MEG data, and in contrast to the model, only the structural effect was decodable. The onset and the peak decodability of the effect varied across constructions. The effect becomes significant first, for the ObjRC-number ($t = 300ms$), then for the PP-Number ($530ms$), and lastly for the PP-animacy ($760ms$). The significance of the decodability was calculated based on cluster-based permutation testing ((51); Methods). The decodability of the structural effect reached its highest value for the ObjRC-number construction ($AUC : 0.61$), followed by the PP-Number ($AUC : 0.58$) and finally by PP-animacy ($AUC : 0.55$).

We observed a discrepancy between the behavioral results, the decoding from the neural-network model, and the decoding from the neural data. At the behavioral level, the structural effect was the dominant factor, and effects emerging from the interaction with the intervening noun were significant only for the number feature. In the model, we observed the same sensitivity across all three main factors, namely, signficant effects arising from the intervening noun for both

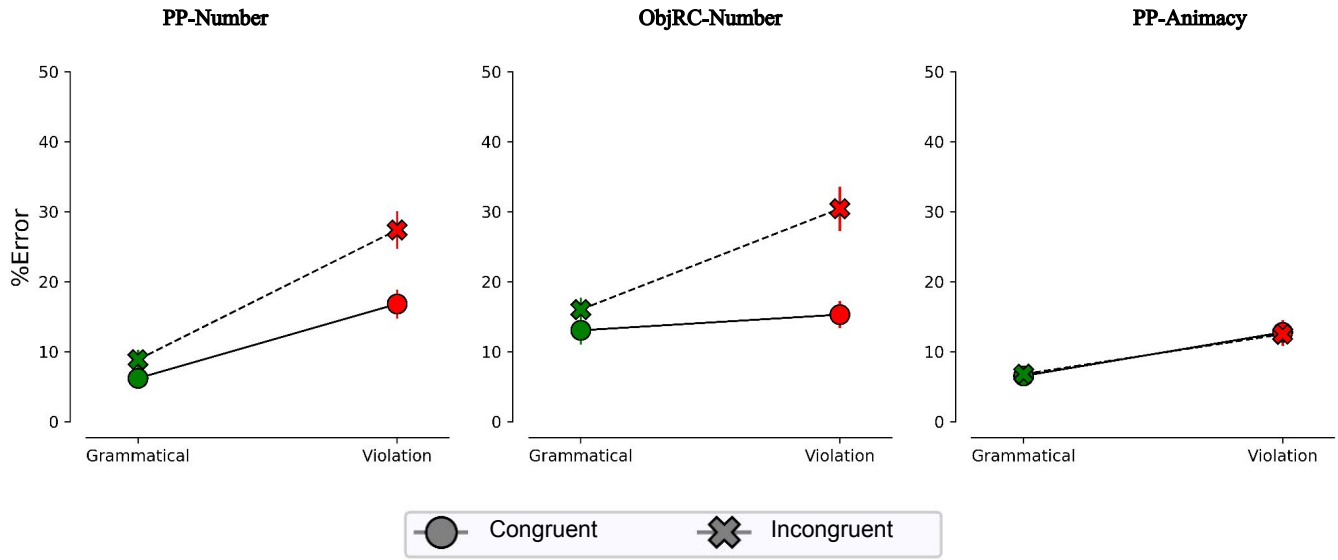


**Fig. 2.** Behavioral results. Interaction plots for the Structural and congruity effects (N=22). The main structural and congruity effects, as well as their interaction, are significant in the number constructions ($p < 0.05$). In the animacy condition, only the structural effect is significant. The error bars indicate the standard error of the mean (SEM) calculated across participants.

the animacy and the number feature. Lastly, in the human neural data, only the structural effect was decodable, whereas the performance of decoding the other effects remained at chance level.

One explanation for the discrepancy between the decoding from the neural data and decoding from the neural model is due to insufficient signal-to-noise ratio (SNR) of the M/EEG data. The simulations we run in the language model might be considered as noiseless "recordings". A similar reason might underlie the discrepancy between the identified effect in the behavioral and neural human data. Given the evidence for the existence of congruity and transition effects from the behavioral analysis, we sought to analyze the data employing a second-order analysis.

**F. The influence of the attractor on the structural effect..** First, we aimed to improve the performance and sensitivity of the decoders by increasing the number of samples on which the decoder is trained. For this, for each construction, we trained a linear binary classifier on the srtuctural effect, and then, separately, tested it on different conditions in the test data (Figure 4A). At training time, the classifier was trained to separate violation and non-violation trials regardless of whether they are congruent or not. Then, at test time, the classifier was separately tested on unseen trials that were either congruent (continuous lines) or incongruent (dashed).

For PP-Number, we found that the structural effect is modulated by congruity. The difference between the congruent and incongruent trials became significant at $580ms$ after the onset of the target, and was sustained for $200ms$. For ObjRC-number, the structural effect was also modulated by congruity. This modulation became significant later compared to the one for the PP-Number construction, starting at $780ms$ and up until $980ms$ after the target onset. Finally, for PP-animacy, the structural effect is not modulated by congruity, in complete agreement with the behavioral results.

**G. The neural correlates of the markedness effect.** In classic work on grammatical agreement, it was observed that for sentences with a long-range subject-verb agreement, participants make more errors if the attractor is plural compared to singular (52). For example, comparing the two following sentences,

1. "The boy near the girls ___".
2. "The boys near the girl ___".

Participants would make on average more errors on (1) compared to (2). This phenomenon is known as the markendess effect, since in English, plural is the marked form (English: Wagers et al. 50, Bock and Miller 52, Eberhard 53 ; Italian: Vigliocco et al. 54 Spanish: Bock et al. 55, Lago et al. 56 French: Franck et al. 57 Russian: Lorimor et al. 58).

Thus far, this phenomenon has been studied mainly through behavioral results. We thus sought to investigate whether we can identify neural correlates of this phenomenon.

We previously saw that the congruity effect modulated the structural effect (Figure 4A). To investigate the markedness effect, we took this analysis further: as before, we trained the classifier on the structural effect, but now we analyzed its performance separately for each value of congruity and attractor number (Figure 4B&C).

When examining the modulation of the structural effect by congruity, we observed no difference between the congruent and incongruent trials in the case of a singular attractor, for none of the three constructions. However, when examining the same effect for the plural attractor, the decodability of the congruent trials was significantly higher compared to the incongruent ones. This difference was significant for the time interval between $620ms$ and up until $810ms$ after the onset of

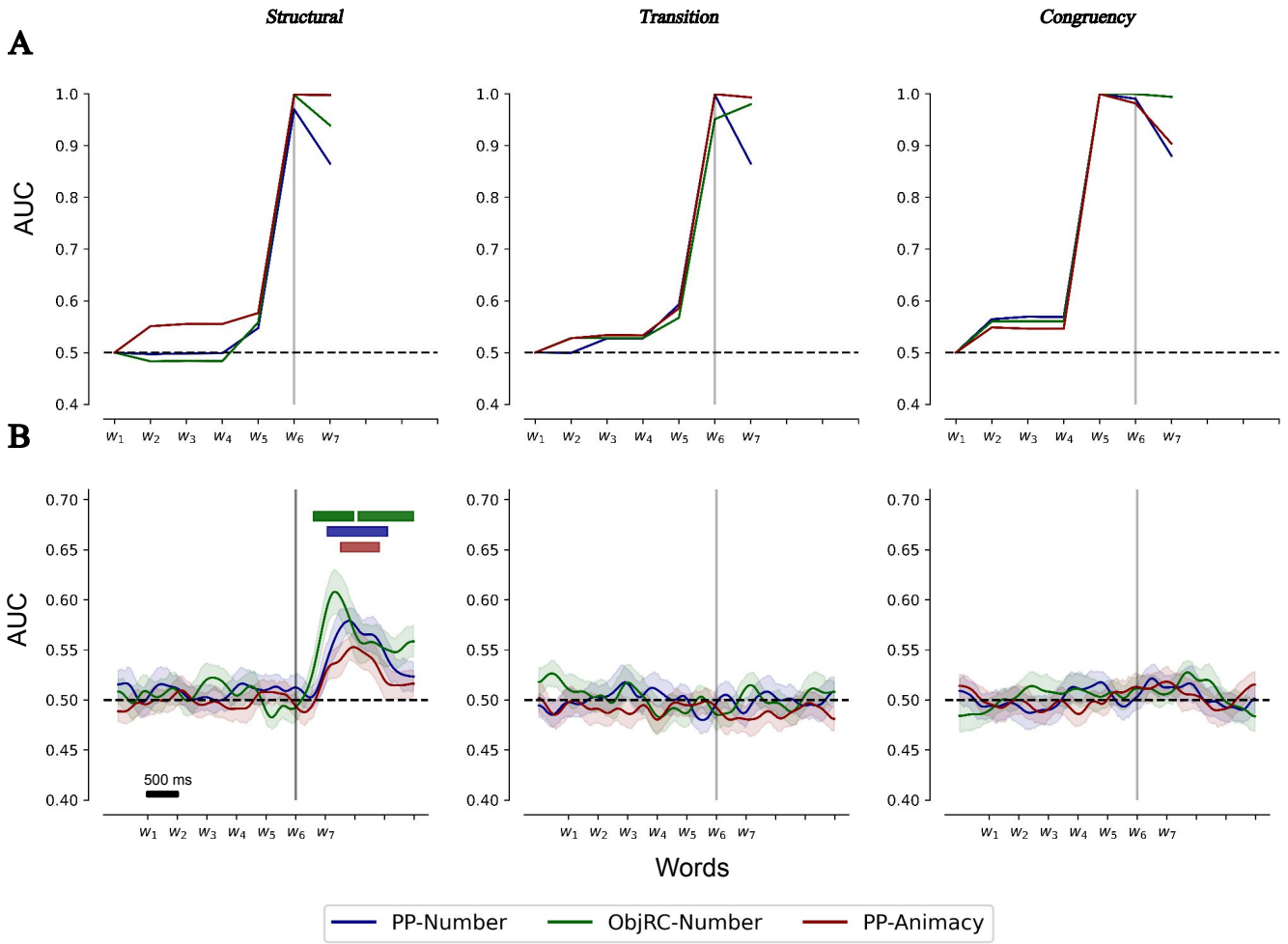


**Fig. 3. Structural but not linear effects are decodable in human data.** (A) Decoding of the main effects originating from the activations of an LSTM architecture. All main effects are decodable. (B) Neural decoding of the three main effects using all sensor types (magnetometers, gradiometers and eeg). A different decoder was trained on each time-point and construction, and evaluated using the Area Under the Curve (AUC). Significant decoding performance was found for the structural effect only. Color stripes indicate statistically significant time intervals ($p < 0.05$; corrected—spatio-temporal clustering permutation test). The decoding for the main effects of transition and congruity remained at chance level until the end of the time window of interest. Results shown for correct responses only, and the data was smoothed with a 100ms moving Gaussian kernel for visualization purpose only.

the target word (3-way interaction cluster-based permutation. $p < 0.05$; Maris and Oostenveld (51))

## Discussion

Evidence from animals, humans and neural language models suggests that non-linguistic sequence processing takes place simultaneously across multiple levels, from local transition-based processing to more global processing that captures long-distance abstract regularities in the input data. Brain-imaging and neurophysiological studies of non-linguistic sequence processing, primarily using the local-global paradigm, have shown that local and global levels of processing are carried by largely distinct neural circuits (31–33, 38, 59). Likewise, analyses of artificial neural networks have concluded that both mechanisms contribute to next-word prediction in AI language models (60). In the present study, we tested whether this is also the case during human language processing: can structure- and transition-based processing be dissociated, and do they have distinct neural underpinning? For this, we introduced a new experimental design which directly contrasts sequential and structure-based processing. The design manipulates prediction violations which are either structure-dependent (i.e., violating expectations that depend on the syntactic structure of the sentence) or structure-agnostic (i.e., violating local word-transition expectations only). We studied subject-verb agreement, and presented sentences with violations of either grammatical number or animacy. This allowed us to test whether structure and transition-based processing are distinguishable, consistently, through the looking glass provided by both types of processing.

The behavioral data revealed a large difference between the two types of violations. Participants were able to detect both number and animacy violations and, in both cases, were better at affirming that a sentence is grammatical/felicitous than at detecting a violation. However, an intervening incongruent noun, with the wrong number, induced a large behavioral interference in the case of number violations, while a similar intervention of a noun with the wrong animacy fea-

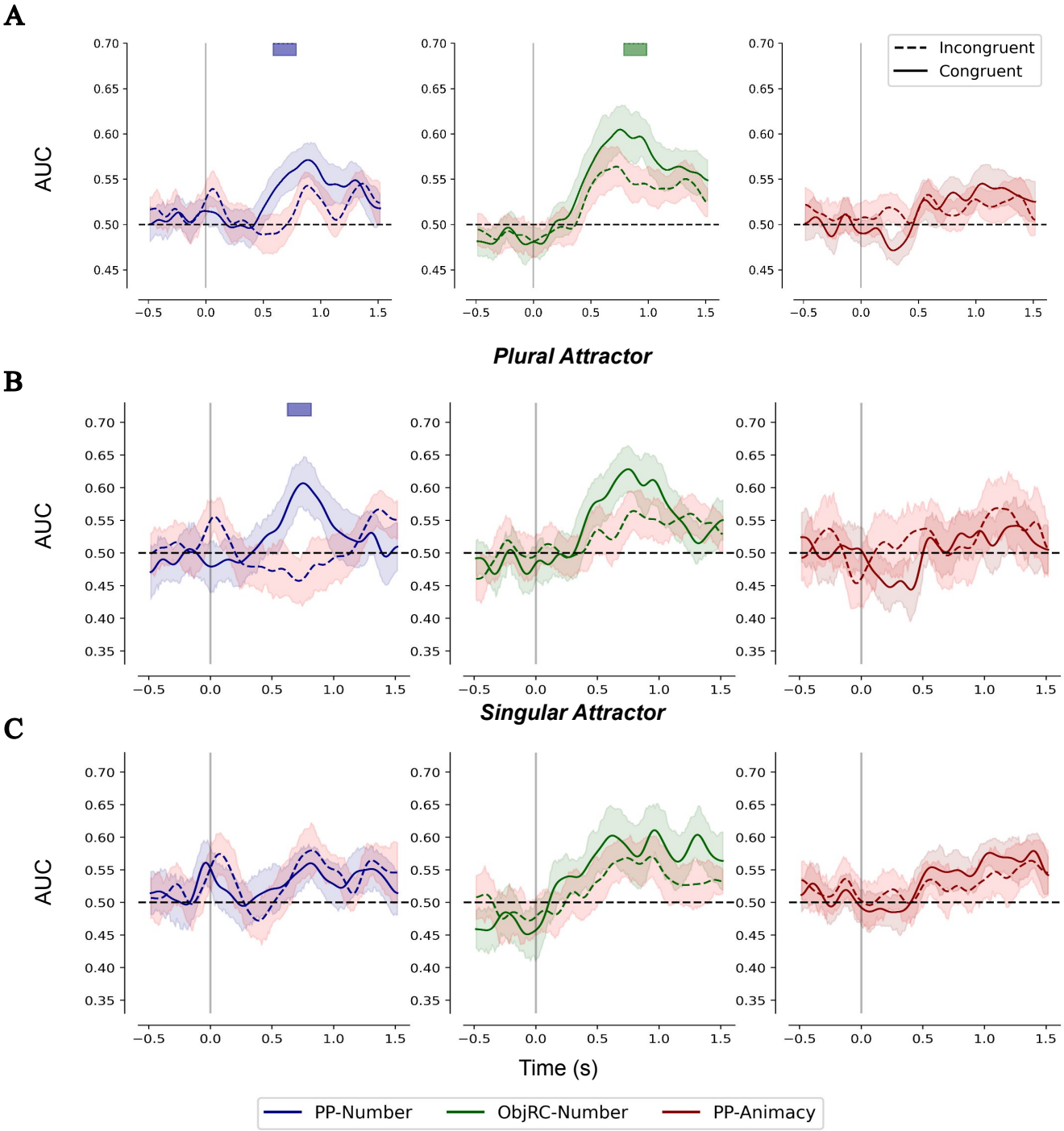


**Fig. 4. Modulation of the structural effect by congruity and neural correlates of the markedness effect.** (A) A classifier was trained on the structural effect (grammatical vs. ungrammatical trials), per construction, and subsequently evaluated on the structural effect when contrasting the congruent (continuous lines) and incongruent (dashed) trials. Color stripes indicate statistically significant time intervals ($p < 0.05$; corrected using spatio-temporal clustering permutation test). congruity was found to modulate the structural effect for both PP-number and ObjRC-number, but not for PP-Animacy. (B&C) For each construction, a classifier was trained on the structural effect, and subsequently tested on the same effect when trials were split for both congruity and attractor number. For PP-number, a statistically significant difference was found for the plural but not the singular case ($p < 0.05$; corrected - cluster based permutation test). For ObjRC-Number and PP-Animacy, no significant differences were found.

ture did not affect participant performance in the case of animacy violations. This finding suggests that those features differed. Although both number and animacy are syntactic as well as semantic features, number is borne by an overt morpheme in both nouns and verbs, whereas animacy information can only be retrieved from the lexicon. Thus, number may have been processed at a morphosyntactic stage earlier than, and more susceptible to interference than, the lexicosemantic stage needed to notice the infelicity of animacy violations. Our results suggest that the latter stage is totally structure-dependent and immune to intervention. At the very least, they indicate that during language processing, grammatical number and animacy are processed and integrated into an ongoing sentence representation in quite different ways, and that the processing of animacy is more robust to intervening material. This result is consistent with memory-based models of sentence processing (61) for which it was suggested that morphosyntactic processing is relatively 'fragile' compared to processing of animacy (62).

Memory-based models of sentence processing (50, 61, 63–67) suggest that new incoming materials, such an inflected verb, trigger memory retrieval of previous information, stored in sentence constituents in memory, in order to complete the noun-verb pairing process. This retrieval process is sensitive to similarities among items in memory, and can therefore explain the observed discrepancy. Morphosyntactic features were found to be weaker cues compared to animacy (62), which results in high similarities among memory items that only differ by morphosyntactic marking. This could make grammatical-number processing more prone to confusion errors, compared to animacy, and therefore to more (erroneous) affirmations of ungrammatical sentences. Our results therefore corroborate the robustness of animacy processing compared to grammatical number.

The second main result from the behavioral data was the identification of transition effects in the case of morphosyntactic processing. In this case, we found a positive interaction between grammaticality and congruity (i.e., a transition effect; Figure 2). That is, an incongruity between the main noun and the local noun elicited more errors in ungrammatical compared to grammatical sentences, more so than in the congruent conditions. This finding is akin to a previous one reported in studies using self-paced reading, which was termed grammatical asymmetry (50, 56). In these studies, it was shown that agreement attraction facilitated the processing of ungrammatical but not grammatical sentences, in the case of incongruent nouns.

Memory-based models of sentence processing (50, 61) were suggested as an explanation for this asymmetry. The cue-based model is a two-stage mechanism. The parser predicts the number of the verb, and only engages in a retrieval process when this prediction mismatches the bottom-up input. The second-stage of this mechanism, is sensitive to interference effects and might lead to the retrieval of the wrong feature.

Returning to the main effect of congruity, the behavioral data is ambiguous and could be explained by two mechanisms. In a sentence such as "the boy near the girls are tall", the incongruent noun could interfere with the decision that the sentence is ungrammatical, either because of the local noun-to-verb transition ("girls are", which is locally correct), or because the incongruent noun-to-noun relationship (the plural of "girls" would interfere with the memory of the singular "boy"). To lift this ambiguity, we looked for the existence of these mechanisms in both human MEG signals and in artificial neural network models. We analyzed those signals using a highly sensitive subject-specific multivariate decoding approach (68), since some of the effects, in particular the transition one, might be small and hard to detect. Decoding methods are adapted for the identification of fine effects in possibly noisy neural data, given that they are multivariate and incorporate model-parameter regularization techniques (69).

In the models, based on previous findings about two neural mechanisms underlying morphosyntactic processing (36), we hypothesized that both the structural and transition effects should be found in the model activations. The structural effect involves a violation between the main noun and verb. Such a violation would affect model predictions about upcoming words once the verb is presented to the model. Regarding the transition effect, our hypothesis was based on previous findings showing that grammatical numbers of the main and local nouns are carried by two separate mechanisms. While the grammatical number of the main noun was shown to be robustly carried by long-range number units up to the verb, the grammatical number of the local noun (attractor) was shown to be carried by short-range number units. Short-range number units encode the grammatical number of the last encountered noun and are agnostic to the syntactic structure of the sentence. Model prediction about upcoming words, and in particular about the main verb, is therefore composed of two (possibility contradicting) predictions, arising from the two types of units in the model. For example, in the PP-number construction, following the transition between the local noun and verb, the encoding in the short-range, but not in long-range number units, changes number. This is expected to be reflected in model activations after verb onset. Indeed, for all three constructions, we found significant decoding of both structural, transition and congruity effects (Figure 3). For the structural and transition effects, maximal decoding was reached at the verb onset, whereas for the congruity effect it occurred on the preceding noun, since information about congruity is already available at this point.

Crucially, the human data contrast with those predictions. For all three constructions, no transition-based effect was observed, and only the structural effect was significant (Figure 3). This suggests that, in the human brain, structure-based computations dominate over transition-based computations during language processing. The results reveal a substantial difference between how non-linguistic sequences and language are processed in humans. Specifically, in the local-global paradigm with sequences of auditory tones, transition effects are easily detectable, while structural effects can be fragile and get reduced, for instance, when subjects are distracted. For language processing, our results differ

drastically: while the structural effect is large and easily detectable, the transition effect is barely detectable or non-existent. Thus, once sequence items enter into the language system, as is the case here for features of number and animacy, sentence-level computations are entirely dominated by structure-sensitive processes, and largely robust to low-level transition effects.

Absence of evidence is not evidence of absence, and the transition-effect might have been undetected due to low signal-to-noise ratio of the MEG data. However, the modulation of the structural effect by congruity (Figure 4A), and further, by grammatical number (Figure 4B&C), shows that our decoding techniques are quite sensitive. This suggests that even if transition-based computations occur during sentence processing, their neural traces are vanishingly small compared to those related to structure. Taken together, our study provides support to the hierarchical view and against the sequential view to sentence-structure processing.

## Methods

A total of 22 participants with normal or corrected to normal vision were included in the M/EEG experiment. In compliance with the institutional guidelines, all participants gave a written, informed consent prior to the experiment and were compensated with 100 Euros for their participation. The participants were native English speakers and prior to the participation, the subjects had to perform an online sentence reading task (Dialang[1] reading & structures task - subjects accepted with a placement greater than C1 in both tests). The procedure and the consent were approved by the local ethical committee (Université Paris-Saclay, ref. CER-Paris-Saclay-2019-063).

**Experimental Paradigm** The participants undertook a rapid serial visual presentation (RSVP) reading task and were asked to report whether a sentence contained a grammatical violation by pressing a button on an MEG response device. To verify that participants understood the task, prior to recording, they went through a short training phase ($10 minutes$). The task was divided into 10 equal runs, where each run contained the same number of trials ($n = 48$). A $600ms$ fixation cross interval preceded the onset of the first word (Figure 1B). All the sentences had the same length. The words were presented with a stimulus onset asynchrony (SOA) of $500ms$. After the onset of the last word and following a time interval of 1s, a decision panel with the words "OK" and "WRONG" appeared on the screen. To control for motor preparation, the location of the words (left or right) was randomized at each trial. As soon as the participants stated their decision, the decision panel disappeared and the subjects received immediate visual feedback on their performance. If their response was correct, they were presented with a green cross, otherwise with a red one. Decision duration was limited to $1.5s$, after which a blue fixation cross appeared and the experiment continued. The interval to the next trial (ITI) was $1.5s$. All time intervals were set to multiples of the video projector refresh rate ($60Hz$).

[1] https://dialangweb.lancaster.ac.uk/

**Stimuli** In this design, we used two linguistic structures, a Prepositional Phrase (PP) and an Object Relative Clause (ObjRC). Additionally, we utilized two features: grammatical number and animacy. This led to the creation of three constructions used in the paradigm: PP-Number, PP-Animacy, ObjRC-Number.

For each of the three constructions (Table 1), we generated 16 stimuli per block for a total of 10 blocks. Half of these stimuli contained a violation. The stimuli were generated using an automated algorithm which sampled without replacement words from the lexicon. Each participant was presented with a different set of stimuli. The lexicon consisted of 19 animate nouns, 7 inanimate nouns and a total of 15 verbs. The stimuli were controlled for low-level features such as length and unigram frequency.

**M/EEG recordings** Due to the COVID-19 pandemic, recording took place in two different MEG centers: NeuroSpin in Saclay, France ($N = 15$), and ICM in Paris ($N = 7$). Participants performed the task while sitting in an electromagnetically shielded room. Brain magnetic fields were recorded with a 306-channel, whole-head MEG by Elekta Neuromag® (Helsinki, Finland), in 102 triplets: one magnetometer and two orthogonal planar gradiometers. In NeuroSpin and ICM, EEG was recorded with a 60 and 64 channel MEG compatible Neuromag EEG cap, respectively. The brain signals were acquired at a sampling rate of $1000Hz$ with a hardware highpass filter at $0.03Hz$. Eye movements and heartbeats were monitored with vertical and horizontal electro-oculograms (EOGs) and electrocardiograms (ECGs). The subjects' head position inside the helmet was measured at the beginning of each run with an isotrack Polhemus Inc. system from the location of four coils placed over frontal and mastoïdian skull areas. All EEG sensors were digitized as well.

**Preprocessing** Bad sensors per sensor-type were automatically detected at the run level based on a variance criterion. Channels of which the variance exceeded the median channel variance by 6 times, or was less than the median variance divided by 6, were marked as bad. A visual inspection was followed to verify the detection accuracy. Prior to the variance detection, Oculomotor and cardiac artifacts were removed at the run level, using signal-space projection (SSP) implemented with MNE Python (70, 71). To compensate for head movement and reduce non-biological noise, the MEG data were Maxwell-filtered (72) using the implementation of Maxwell filtering in MNE Python. The bad EEG sensors were interpolated using the spherical spline method (73) implemented in the same package. Following Maxwell filtering, the linear component of the data was removed and the time-series were clipped at the upper and lower bound values of (-3,3) interquartile range (IQR) around the median. The data were then bandpass filtered between 0.4 and 50 Hz using a linear-phase FIR filter (hamming) with delay compensation, implemented in MNE-python version 0.16 (70). Finally, the

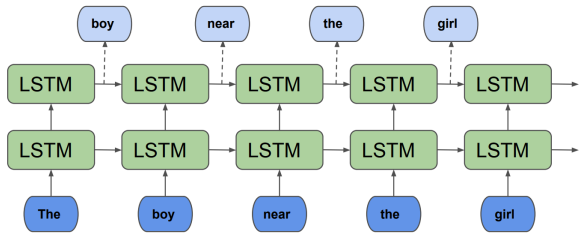


**Fig. 5.** Graphical description of a two-layer recurrent neural language model with LSTM cells (not discussed here; see, e.g., 75). At each timestep, the model processes an input word and outputs a probability distribution over potential next words in the sentence. The prediction of the output word depends on both the input word and on the previous state of the model, which serves as longer-term context (the horizontal arrows in the figure represent the *recurrent* connections carrying the previous state through).

continuous time-series were segmented into 3.5s epochs of interest (first word onset to panel onset) and the SSP procedure was applied to the epoched data to remove heart-beats and ocular motions.

**Decoding Analyses** We used a temporally defined decoding approach to classify neural activity from two conditions at the trial level (34, 59). These analyses were implemented in MNE-python version 0.16 (70). Prior to model fitting, the data were standardized using the Scikit-Learn package (74). We used a linear classifier (logistic-regression) with default Scikit-Learn parameters. The evaluation metric was the Area Under the Curve (AUC). The estimator was trained and tested on data from the same condition. To prevent overfitting, we used a 5-fold stratified cross-validation procedure.

**Statistical Analyses** The reported statistics correspond to group-level analyses and were performed using the Statsmodels package in Python3 and in MNE-python version 0.16 (70). The statistical significance of the decoding performance over time was evaluated and corrected for multiple comparisons using a cluster-based permutation approach (51), using a total of 1000 permutations. The significance threshold (alpha level) for all analyses was set to 0.05.

**Neural Language Models** In the computational experiments, we use a Neural Language Model (NLM). A NLM is a language model implemented by a neural network, defining a probability distribution over sequences of words. It factorizes the probability of a sentence into a multiplication of the conditional probabilities of all words in the sentence, given the words that precede them:

$$P(w_1 w_2 \cdots w_n) = \prod_i p(w_i | w_1 \cdots w_{i-1}) \qquad \textbf{(1)}$$

This type of language model can thus be used as *next-word predictor*: given the preamble of a sentence, it outputs a probability distribution over potential next words. We exploit this fact in our experiments.

**Model description** The specific NLM we use in our experiments is the English NLM made available by Gulordava et al. (76)[2]. It is a recurrent LSTM language model (77), consisting of two layers, each with 650 Long-Short Term Memory units (78), input and output embedding layers of 650 units and input and output layers of size 50000 (the size of the vocabulary). The weights of the input and output embedding layers are not shared (79). The last layer of the model is a softmax layer, whose activations sum up to 1 and as such corresponds to a probability distribution over all words in the NLM's vocabulary.

**Model training** The weights of a NLM are typically tuned by presenting them with large amounts of data (a *training corpus*) and providing them feedback on how well they can predict each next word in the running text. This allows them to adjust their parameters to maximize the probabilities of the sentences in the corpus. Our NLM was trained on a sample of the English Wikipedia text, containing 100M word tokens and 50K word types. Further details can be found in Gulordava et al. (76).

**ACKNOWLEDGEMENTS**
The authors would like to thank Christophe Pallier, Théo Desbordes, Jean-Rémi King and Luigi Rizzi for interesting discussions and feedback on this work. We are also grateful to Fosca Al Roumi, Leila Azizi, as well as the MEG team of ICM for helping with the data acquition. This work was supported by INSERM, CEA, Collège de France, the Bettencourt-Schueller Foundation and an ERC grant, "NeuroSyntax" project, to S.D.

## Bibliography

1. Noam Chomsky. *Syntactic structures*. Mouton publishers, 1957. ISBN 978-90-279-3385-0. Accepted: 2019-04-10T06:25:21Z Journal Abbreviation: Janua linguarum.
2. Marc Hauser, Noam Chomsky, and Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579, 11 2002.
3. Martin BH Everaert, Marinus AC Huybregts, Noam Chomsky, Robert C Berwick, and Johan J Bolhuis. Structures, not strings: linguistics as part of the cognitive sciences. *Trends in cognitive sciences*, 19(12):729–743, 2015.
4. Stefan Frank, Rens Bod, and Morten Christiansen. How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, 279(1747):4522–4531, 2012.
5. Ewa Dąbrowska and Elena Lieven. Towards a lexically specific grammar of children's question constructions. 2005.
6. Ewa Dąbrowska. What exactly is universal grammar, and has anyone seen it? *Frontiers in psychology*, 6:852, 2015.
7. Noam Chomsky. *Lectures on government and binding: The Pisa lectures*. Number 9. Walter de Gruyter, 1993.
8. Noam Chomsky, Peter W Culicover, Thomas Wasow, Adrian Akmajian, et al. On wh-movement. *1977*, 65, 1977.
9. Mark Liberman and Alan Prince. On stress and linguistic rhythm. *Linguistic inquiry*, 8(2): 249–336, 1977.
10. Noam Chomsky. *Essays on Form and Interpretation*. North Holland, New York, 1977.
11. Joseph Aoun and David W Lightfoot. Government and contraction. *Linguistic Inquiry*, 15 (3):465–473, 1984.
12. Noam Chomsky. Problems of knowledge and freedom: The russell lectures. 1971.
13. Massimo Piattelli-Palmarini. Language and learning: the debate between jean piaget and noam chomsky. 1980.
14. W Tecumseh Fitch and Marc D Hauser. Computational constraints on syntactic processing in a nonhuman primate. *Science*, 303(5656):377–380, 2004.
15. Mariacristina Musso, Andrea Moro, Volkmar Glauche, Michel Rijntjes, Jürgen Reichenbach, Christian Büchel, and Cornelius Weiller. Broca's area and the language instinct. *Nature neuroscience*, 6(7):774–781, 2003.
16. Michal Ben-Shachar, Dafna Palti, and Yosef Grodzinsky. Neural correlates of syntactic movement: converging evidence from two fmri experiments. *Neuroimage*, 21(4):1320–1336, 2004.
17. Angela D Friederici, Jörg Bahlmann, Stefan Heim, Ricarda I Schubotz, and Alfred Anwander. The brain differentiates human and non-human grammars: functional localization and structural connectivity. *Proceedings of the National Academy of Sciences*, 103(7):2458–2463, 2006.
18. Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392, 2011.
19. Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, 2011.

[2] https://github.com/facebookresearch/colorlessgreenRNNs

20. Einat Shetreet and Naama Friedmann. The processing of different syntactic structures: fmri investigation of the linguistic distinction between wh-movement and verb movement. *Journal of Neurolinguistics*, 27(1):1–17, 2014.
21. Emiliano Zaccarella and Angela D Friederici. Merge in the human brain: A sub-region based functional investigation in the left pars opercularis. *Frontiers in psychology*, 6:1818, 2015.
22. Matthew Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney Cash, Lionel Naccache, John Hale, Christophe Pallier, and Stanislas Dehaene. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678, 2017.
23. Stewart M McCauley and Morten H Christiansen. Learning simple statistics for language comprehension and production: The cappuccino model. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.
24. Morten H Christiansen, M Louise Kelly, Richard C Shillcock, and Katie Greenfield. Impaired artificial grammar learning in agrammatism. *Cognition*, 116(3):382–393, 2010.
25. Jennifer B Misyak and Morten H Christiansen. When 'more'in statistical learning means 'less' in language: individual differences in predictive processing of adjacent dependencies. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, 2010.
26. Maureen Gillespie and Neal J Pearlmutter. Hierarchy and scope of planning in subject–verb agreement production. *Cognition*, 118(3):377–397, 2011.
27. Maureen Gillespie and Neal J Pearlmutter. Against structural constraints in subject–verb agreement production. *Journal of experimental psychology: Learning, memory, and cognition*, 39(2):515, 2013.
28. Christopher M Conway and David B Pisoni. Neurocognitive basis of implicit learning of sequential structure and its relation to language processing. *Annals of the New York Academy of Sciences*, 1145(1):113–131, 2008.
29. Karl-Magnus Petersson, Vasiliki Folia, and Peter Hagoort. What artificial grammar learning reveals about the neurobiology of syntax. *Brain and language*, 120(2):83–95, 2012.
30. Morten H Christiansen, Christopher M Conway, and Luca Onnis. Similar neural correlates for language and sequential learning: Evidence from event-related brain potentials. *Language and cognitive processes*, 27(2):231–256, 2012.
31. Tristan A Bekinschtein, Stanislas Dehaene, Benjamin Rohaut, François Tadel, Laurent Cohen, and Lionel Naccache. Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences*, 106(5):1672–1677, 2009.
32. Imen El Karoui, Jean-Remi King, Jacobo Sitt, Florent Meyniel, Simon Van Gaal, Dominique Hasboun, Claude Adam, Vincent Navarro, Michel Baulac, Stanislas Dehaene, et al. Event-related potential, time-frequency, and functional connectivity facets of local and global auditory novelty processing: an intracranial study in humans. *Cerebral cortex*, 25(11):4203–4212, 2015.
33. Zenas C Chao, Kana Takaura, Liping Wang, Naotaka Fujii, and Stanislas Dehaene. Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. *Neuron*, 100(5):1252–1266, 2018.
34. Jean-Rémi King and Stanislas Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, 18(4):203–210, 2014.
35. Melanie Strauss, Jacobo D Sitt, Jean-Remi King, Maxime Elbaz, Leila Azizi, Marco Buiatti, Lionel Naccache, Virginie Van Wassenhove, and Stanislas Dehaene. Disruption of hierarchical predictive coding during sleep. *Proceedings of the National Academy of Sciences*, 112(11):E1353–E1362, 2015.
36. Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in lstm language models, 2019.
37. Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699, 2021. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2021.104699.
38. Stanislas Dehaene, Florent Meyniel, Catherine Wacongne, Liping Wang, and Christophe Pallier. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron*, 88(1):2–19, 2015.
39. Felipe Pegado, Tristan Bekinschtein, Nicolas Chausson, Stanislas Dehaene, Laurent Cohen, and Lionel Naccache. Probing the lifetimes of auditory novelty detection processes. *Neuropsychologia*, 48(10):3145–3154, 2010.
40. Lynn Uhrig, Stanislas Dehaene, and Béchir Jarraya. A hierarchy of responses to auditory regularities in the macaque brain. *Journal of Neuroscience*, 34(4):1127–1132, 2014.
41. Julie Franck, Ulrich Hans Frauenfelder, and Luigi Rizzi. A syntactic analysis of interference in subject–verb agreement. *MIT working papers in linguistics*, (53):173–190, 2007.
42. Lee Osterhout and Phillip J Holcomb. Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, 31(6):785–806, 1992.
43. Lee Osterhout and Linda A Mobley. Event-related brain potentials elicited by failure to agree. *Journal of Memory and language*, 34(6):739–773, 1995.
44. Karl Friston. A theory of cortical responses. 360(1456):815–836, April 2005. doi: 10.1098/rstb.2005.1622.
45. Gina R Kuperberg and T Florian Jaeger. What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59, 2016.
46. Karl Friston, Rosalyn J. Moran, Yukie Nagai, Tadahiro Taniguchi, Hiroaki Gomi, and Josh Tenenbaum. World model learning and inference. 144:573–590, December 2021. doi: 10.1016/j.neunet.2021.09.011.
47. Angela D Friederici, Karsten Steinhauer, and Stefan Frisch. Lexical integration: Sequential effects of syntactic and semantic information. *Memory & cognition*, 27(3):438–453, 1999.
48. Peter Hagoort. Interplay between syntax and semantics during sentence comprehension: Erp effects of combining syntactic and semantic violations. *Journal of cognitive neuroscience*, 15(6):883–899, 2003.
49. Colin Phillips, Matthew W Wagers, and Ellen F Lau. Grammatical illusions and selective fallibility in real-time language comprehension. *Experiments at the Interfaces*, 37:147–180, 2011.
50. Matthew W Wagers, Ellen F Lau, and Colin Phillips. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237, 2009.
51. Eric Maris and Robert Oostenveld. Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190, 2007.
52. Kathryn Bock and Carol A Miller. Broken agreement. *Cognitive psychology*, 23(1):45–93, 1991.
53. Kathleen M Eberhard. The marked effect of number on subject–verb agreement. *Journal of Memory and language*, 36(2):147–164, 1997.
54. Gabriella Vigliocco, Brian Butterworth, and Carlo Semenza. Constructing subject-verb agreement in speech: The role of semantic and morphological factors. *Journal of Memory and Language*, 34(2):186–215, 1995.
55. Kathryn Bock, Manuel Carreiras, and Enrique Meseguer. Number meaning and number grammar in english and spanish. *Journal of Memory and Language*, 66(1):17–37, 2012.
56. Sol Lago, Diego E Shalom, Mariano Sigman, Ellen F Lau, and Colin Phillips. Agreement attraction in spanish comprehension. *Journal of Memory and Language*, 82:133–149, 2015.
57. Julie Franck, Gabriella Vigliocco, and Janet Nicol. Subject-verb agreement errors in french and english: The role of syntactic hierarchy. *Language and cognitive processes*, 17(4): 371–404, 2002.
58. Heidi Lorimor, Kathryn Bock, Ekaterina Zalkind, Alina Sheyman, and Robert Beard. Agreement and attraction in russian. *Language and cognitive processes*, 23(6):769–799, 2008.
59. Stanislas Dehaene and Jean-Rémi King. Decoding the dynamics of conscious perception: The temporal generalization method. *Micro-, meso-and macro-dynamics of the brain*, pages 85–97, 2016.
60. Yair Lakretz, Stanislas Dehaene, and Jean-Rémi King. What limits our capacity to process nested long-range dependencies in sentence comprehension? *Entropy*, 22(4):446, 2020.
61. Richard L Lewis and Shravan Vasishth. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29(3):375–419, 2005.
62. Anastasia Stoops and Kiel Christianson. Parafoveal processing of inflectional morphology on russian nouns. *Journal of Cognitive Psychology*, 29(6):653–669, 2017.
63. Brian McElree, Stephani Foraker, and Lisbeth Dyer. Memory structures that subserve sentence comprehension. *Journal of memory and language*, 48(1):67–91, 2003.
64. Julie A Van Dyke and Clinton L Johns. Memory interference as a determinant of language comprehension. *Language and linguistics compass*, 6(4):193–211, 2012.
65. William Badecker and Frantisek Kuminiak. Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in slovak. *Journal of memory and language*, 56(1):65–85, 2007.
66. Brian Dillon, Alan Mishler, Shayne Sloggett, and Colin Phillips. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103, 2013.
67. Andrea E Martin and Brian McElree. A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, 58(3):879–906, 2008.
68. Jean-Rémi King and Stanislas Dehaene. Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210, 2014.
69. Jean-Rémi King, Laura Gwilliams, Chris Holdgraf, Jona Sassenhagen, Alexandre Barachant, Denis Engemann, Eric Larson, and Alexandre Gramfort. Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition. 2018.
70. Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7:267, 2013.
71. Mainak Jas, Eric Larson, Denis A Engemann, Jaakko Leppäkangas, Samu Taulu, Matti Hämäläinen, and Alexandre Gramfort. A reproducible meg/eeg group study with the mne software: recommendations, quality assessments, and good practices. *Frontiers in neuroscience*, 12:530, 2018.
72. Samu Taulu, Matti Kajola, and Juha Simola. Suppression of interference and artifacts by the signal space separation method. *Brain topography*, 16(4):269–275, 2004.
73. François Perrin, Jacques Pernier, O Bertrand, and Jean Francois Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, 72(2):184–187, 1989.
74. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12: 2825–2830, 2011.
75. Yoav Goldberg. *Neural Network Methods for Natural Language Processing*. Morgan & Claypool, San Francisco, CA, 2017.
76. Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL*, pages 1195–1205, New Orleans, LA, 2018.
77. Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, Berlin, 2012.
78. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
79. Ofir Press and Lior Wolf. Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859*, 2016.